Barigozzi, Matteo and Moneta, Alessio

Identifying the independent sources of consumption variation

Article (Accepted version)
(Refereed)

IDENTIFYING THE INDEPENDENT SOURCES OF CONSUMPTION VARIATION

Matteo BARIGOZZI[1] Alessio MONETA[2]

September 24, 2014

Abstract

By representing a system of budget shares as an approximate factor model we determine its rank, i.e. the number of common functional forms, or factors and we estimate a base of the factor space by means of approximate principal components. We assume that the extracted factors span the same space of basic Engel curves representing the fundamental forces driving consumers' behaviour. We identify these curves by imposing statistical independence and by studying their dependence on total expenditure using local linear regressions. We prove consistency of the estimates. Using data from the U.K. Family Expenditure Survey from 1977 to 2006, we find strong evidence of two common factors and mixed evidence of a third factor. These are identified as decreasing, increasing, and almost constant Engel curves. The household consumption behaviour is therefore driven by two factors respectively related to necessities (e.g. food), luxuries (e.g. vehicles), and in some cases by a third factor related to goods to which is allocated the same percentage of total budget both by rich and poor households (e.g. housing).

Keywords: Budget Shares; Engel Curves; Approximate Factor Models; Independent Component Analysis; Local Linear Regression.

JEL classification: C52, D12.

[1] Department of Statistics, London School of Economics and Political Science, Houghton Street, WC2A 2AE, London, United Kingdom. Email: *m.barigozzi@lse.ac.uk*

[2] Institute of Economics, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà 33, 56127 Pisa, Italy. Email: *amoneta@sssup.it*

We thank two anonymous referees for their helpful comments. We also thank Stephan Bruns, Andreas Chai, Giorgio Fagiolo, Mark Trede, and Ulrich Witt for their comments, and Stefanie Picard for research assistance. Finally, we thank the U.K. Central Statistical Office for making available the U.K. Family Expenditure Survey and the Expenditure and Food Survey data through the Economic and Social Data Service.

1 Introduction

In his seminal work of 1857, Ernst Engel made already clear that all kinds of household expenditures depend on income, but each type of expenditure depends on income in its own way. The functional dependence of expenditure on income is traditionally studied by the analysis of Engel curves. These are regression functions in which the dependent variable is the level or share of expenses (i.e. the budget share) allocated towards a category of goods or services and the explanatory variable is income, usually proxied by total expenditure. Typically, Engel curves estimated over different samples of households show that budget shares change with income, which implies that for many types of expenditures the levels grow non-proportionally with income. For example, the total budget allocated on food tends to decrease with income. This is a very robust empirical regularity, found in numerous samples of families, and classically referred to as *Engel law*. Other types of expenditure follow different patterns, although in a less robust manner. For example, it is often the case to observe budget shares spent on leisure goods or services which increase with income.

The various reactions to income changes, showed by different types of expenditures, suggest the existence of different motives driving consumption decisions. Each motive determines a very specific reaction to income changes and all observed Engel curves are to be interpreted as a mixture of these basic reactions. This paper presents a statistical analysis of the variety of expenditure patterns (across some categories of goods and services) with the aim of capturing the (unobserved) reactions to income changes caused by the underlying motives.

The literature trying to interpret the various shapes of Engel curves in terms of underlying motives traces back to Ernst Engel (1857) himself. He suggested that when studying household consumption we should distinguish and classify expenditure categories according to the wants they serve (see Chai and Moneta, 2010). He identified particular categories of wants as “nourishment”, “clothing”, “housing”, “recreation”, “safety”, and several others. To each category of expenditure it should be assigned one want or an homogeneous set of wants. In this framework, the shape of the Engel curve for food (that is the Engel law) can be explained by asserting that nourishment is one of the basic human needs and that the goods which are necessary for their satisfaction have, in case of deprivation, higher utility than that of any other commodities. Yet, once the want for nourishment is satiated, the marginal utility of successive increments of the same goods falls (see Pasinetti, 1981; Witt, 2001). Thus, each family seeks to reach a certain level of expenses on food (under the constraint of its budget), but once its members are nourished enough, other types of expenditures will be considered, if there is enough budget left. This would explain why poor families spend, on average, a higher share of their budget on food than rich families. Other assumptions on the relationship between single wants and utility and on the existence of a hierarchy of wants may help explain the structure of Engel curves for higher order goods and services, included luxuries (see Pasinetti, 1981; Foellmi and Zweimüller, 2008).

It is, however, very problematic to assign to each category of expenditure an homogeneous set of motives. Food expenditure and consumption may well be predominantly driven by need of calories intake, which is genetically determined and therefore shared (with the usual genetic variance) among all humans (see Witt, 1999). But other motives, of very different nature, may concur in influencing the decision about the budget share to be allocated on food, like, for example, the need of social recognition, health, etc. Categories like clothing, housing, leisure goods and services, travel, etc. appear even more problematic to be assigned to a class of homogeneous wants. Travel expenditures, for instance, may be driven by very different kinds of motives, like leisure, health and social recognition. Moreover, the existence of a hierarchy of wants is empirically controversial (see Banerjee and Duflo, 2011).

In this paper we assume that there are different motives driving consumption decisions. We conjecture that each of these motives determines a specific reaction to income changes and we estimate and identify each of these reactions, which are interpreted as *basic* (latent) independent Engel curves. The assumption of statistical independence is grounded on an argument about the specific nature of each of the underlying motives. The observed Engel curves are then modelled as mixtures, i.e. linear combinations, of the basic curves. This means that in each category of expenditure all motives can in principle concur in driving the reaction of consumption to the income–stimulus. By means of factor analysis, combined with independent component analysis, we estimate and identify the shape of the basic Engel curves, their number, and the coefficients of the linear combinations that give rise to the observed Engel curves.

Following Lewbel (1991), we consider a system of budget shares that are linearly driven by few latent variables, which in turn are functions of total expenditure. This system can be viewed as a latent factor model for the observed budget shares. We estimate, in particular, an approximate factor model, which allows idiosyncratic terms to be mildly correlated. These models deal with panel of data which are large in both dimensions (number of variables and observations). In this manner, they overcome the problem of non-zero correlation among idiosyncratic terms (see e.g. Stock and Watson, 1989; Forni et al., 2000; Bai and Ng, 2002; Doz et al., 2012, 2011, among others).

We use deflated expenditure data of the U.K. Family Expenditure Survey and the Expenditure and Food Survey relative to 13 expenditure categories and based on surveys conducted on different households between 1977 and 2006. In order to estimate an approximate factor model, we need to build a large panel, in terms of both the number of types of budget shares (expenditure categories) and the dimension over which the same budget shares are repeatedly observed. This second dimension is not, in our case, time, as in the typical factor-model setting (1977–2006 would be a too short time series), but total expenditure. We obtain a panel with large dimensions by pooling the budget shares relative to the 13 categories over different 10 years windows. The second dimension does not consist of time points but of 100, income determined, representative households. Depending on the number of household members, we consider different datasets built in this way and on each of them the analysis is repeated. This approach is similar to Kneip (1994) and further technical details and and empirical justifications are given below.

Exploiting this large dataset, we determine the number of basic Engel curves, i.e. the rank of the system, using the criteria for the number of common factors by Bai and Ng (2002) and by Alessi et al. (2010), and the test by Onatski (2010). We then estimate the factors by means of principal component analysis. The determination of the rank of systems of Engel curves has concerned much literature on empirical analysis of consumption (see Gorman, 1981; Lewbel, 1991; Kneip, 1994; Donald, 1997; Banks et al., 1997, among others).

Since factor analysis is not sufficient to identify the latent Engel curves, we need to apply an additional technique which allows us to study their functional form. This technique, referred to as independent component analysis (see Comon, 1994; Hyvärinen et al., 2001), exploits the observed non–Gaussianity of the estimated factors and the assumption of statistical independence of the basic Engel curves, in order to obtain the appropriate orthogonal transformation of estimated factors. Having identified the correct factors, we investigate what kind of functional dependencies on total expenditure they convey. These functional dependencies are the basic Engel curves, which we estimate and interpret by means of parametric and non–parametric methods (see Lewbel, 1991, for the parametric approach).

In our data we find clear evidence of two common factors driving the household consumption choices with in some case also a third factor playing a role. The first two factors correspond to different functions of total expenditure related to the standard classification of goods: *i*) a decreasing function capturing consumption necessities (e.g. food), *ii*) an increasing function related to luxuries (e.g. vehicles). Finally, the third factor when present is associated to an almost constant function corresponding to the expenditure for goods to which is allocated the same percentage of total budget both in rich and in poor households (e.g. housing).

In section 2, we outline the budget shares model considered in this paper. In sections 3 and 4, we describe the way in which we build the dataset and we give empirical motivations for the assumptions made. In section 5, we represent the system as an approximate factor model, we explain the approximate principal components estimation method, the related criterion for the number of common factors, and the identification via independent component analysis. In section 6, we give two consistency results for the estimated basic Engel curves. In section 7, we show results on the number of factors and their interpretation as non–linear functions of total expenditure. Finally, in section 7, we conclude. Data description, and additional results related to other samples not considered in the paper are available in a complementary appendix available on–line from the authors webpages.

2 The model for budget shares

A system of Engel curves describes how expenditures on a set of categories of goods and services change as the household's budget increases in a particular price regime, i.e. holding prices fixed. Let w_{gh} be the budget share of a category of goods or services g that the household h buys. Considering G categories of expenditure and a sample of H households, holding prices fixed, a system of Engel curves can be written as

$$w_{gh} = m_g(x_h) + e_{gh}, \quad g = 1, \dots, G, \;\; h = 1, \dots, H,$$

where x_h is total expenditure (income for short). The term $m_g(x_h)$ describes the dependence of each budget share on the total budget. It is a regression function (conditional expectation function), while e_{gh} is an independent error term. Thus, $m_g(x_h)$ can be directly estimated with parametric or non–parametric methods. However, based on the idea of basic Engel curves driving the observed household behaviour, we write each observed Engel curve as a linear combination of $R < G$ latent independent Engel curves:

$$w_{gh} = \sum_{r=1}^{R} a_{gr} f_r(x_h) + e_{gh} = \mathbf{a}_g \mathbf{f}(x_h) + \; e_{gh}, \quad g = 1, \dots, G, \;\; h = 1, \dots, H. \qquad (1)$$

In this framework, R is the rank of the matrix $\mathbf{A} = (\mathbf{a}_1' \ldots \mathbf{a}_G')'$ and it determines the dimension of the space spanned by the basic Engel curves $f_1(x_h), \ldots, f_R(x_h)$. Gorman (1981) and Lewbel (1991) prove that the knowledge of R can provide us with important implications about the functional form, separability, and aggregability of consumer preferences. In particular, if $R = 1$ and the *adding–up* condition holds, then budget shares are constant across income. If $R = 2$, then the underlying demand functions are generalized linear. For example the Almost Ideal and Translog models of Deaton and Muellbauer (1980) and Jorgensen and Stoker (1982) are rank-two models. If the system of equations (1) is an *exactly aggregable* class of demand, and if the underlying utility functions are restricted to be consistent with the exactly aggregable class, then utility maximization requires $R \leq 3$.[1]

In general, however, utility maximization does not require demand systems to have $R \leq 3$, nor the finding of $R \leq 3$ implies the presence utility–maximizer consumers.[2] Indeed, Aversi et al. (1999) simulate micro–founded models of consumption expenditure which generate rank-three systems of demand despite the fact that the simulated individual behaviours are designed by the authors to be at odds with those postulated by the standard utility–based model of rational choice.

If we could observe household expenditures over different time periods, the underlying system of Engel curves, (1), would become

$$w^*_{gh\,t} = \sum_{r=1}^{R_t} a_{gr\,t} f_{r\,t}(x^*_{h\,t}, \mathbf{p}_t) + e_{gh\,t}, \quad g = 1, \ldots, G, \;\; h = 1, \ldots, H, \;\; t = 1, \ldots, T. \qquad (2)$$

Since we deal with different time periods, we assume that at each point in time t, there is a particular price regime, determined by a vector of prices $\mathbf{p}_t$ and we denote by $x^*_{h\,t}$ nominal total expenditure and by $w^*_{gh\,t}$ nominal budget shares, i.e. the ratio of a nominal expenditure on $x^*_{h\,t}$. Notice that in principle in (2) also the number of basic Engel curves, R_t, could change with time. A possible specification for model (2) is based on deflated data:

$$w^*_{gh\,t}\frac{\bar{p}_t}{p_{gt}} = \sum_{r=1}^{R_t} a_{gr\,t} f_{r\,t}\left(\frac{x^*_{h\,t}}{\bar{p}_t}\right) + e_{gh\,t}, \quad g = 1, \ldots, G, \;\; h = 1, \ldots, H, \;\; t = 1, \ldots, T, \qquad (3)$$

where $\bar{p}_t$ is the aggregate price index, and p_{gt} is the price index for the category of expenditure g. We deflate both the budget share and nominal total expenditure by a price index. The budget share, being equal to a ratio where the numerator is the (nominal) level of expenditure and the

[1] In an exactly aggregable class of demand the aggregate (across households) demand depends only on the means of the individual demands and individual heterogeneity can be neglected. For a discussion see Kirman (1992), Stoker (1993), Hildenbrand (2008)

[2] We thank one referee for having clarified this.

denominator is the (nominal) total budget, is divided by a price index for the particular expenditure and multiplied (in order to deflate the denominator) by the total price index. Total expenditure on the right hand side is deflated by dividing it by the total price index. Hereafter, real budget shares and real total expenditure are denoted as $w_{gh\,t}$ and $x_{h\,t}$ respectively. Model (3) belongs to the class of Deflated Income Demand models which have been studied by Lewbel (2003), who shows that they can have rank four without violating the hypothesis of utility maximization.

We could estimate (3) only if we had a balanced panel of budget shares, but this is not possible with the data at hand as we cannot monitor the same household across different years. Thus we modify (3) by intervals of total expenditure which define representative households that keep total expenditure constant over time. Technical details are given in the next section. In this way, we can pool the budget shares of representative households across years and further simplify the model. In particular, by letting $j = 1, \ldots, J$ correspond to the pooled categories of expenditures over time with $J = GT$, we consider the model

$$w_{jh} = \sum_{r=1}^{R} a_{jr} f_r(x_h) + e_{jh}, \quad j = 1, \ldots, J, \;\; h = 1, \ldots, H. \tag{4}$$

When comparing model (4) with (3), it has to be noticed that the coefficients a_{jr} are still taking into account changes over time. Indeed, if $G = 13$ (as in our empirical application) then, for example, $a_{1\,r}$ and $a_{14\,r}$ correspond to observations of the budget shares $w_{1\,h}$, $w_{14\,h}$ which refer to the same category of expenditure ($g = 1$) but observed at different points of time ($t = 1$ and $t = 2$). On the other hand model (4) is based on three new assumptions: *i*) the distribution of real total expenditure is stable across the considered years; *ii*) the structure and number of latent Engel curves is not changing over time; *iii*) there exist representative households, which implies that in given intervals of total expenditure demands are exactly aggregable. While the first two hypothesis are empirically justified in section 4, the latter can be considered reasonable if the chosen intervals are small enough. The question as to what restrictions such hypothesis imposes on the rank of the system of Engel curves is clearly of interest but we do not investigate it here. We just notice that model (4) is closely related to the Deflated Income Demand model.

We then follow another direction of research and focus on the estimation of the functional form of the basic latent Engel curves. As suggested by Bai and Ng (2002), we consider the R basic Engel curves as common factors which in turn can be thought as non–linear functions of total expenditure $g_r(x_h)$ plus an error term z_{rh}. Thus the model we empirically estimate is

$$\begin{aligned} w_{jh} &= \sum_{r=1}^{R} a_{jr} f_r(x_h) + e_{jh} = && (5) \\ &= \sum_{r=1}^{R} a_{jr} g_r(x_h) + \sum_{r=1}^{R} a_{jr} z_{rh} + e_{jh}, \quad j = 1, \ldots, J, \quad h = 1, \ldots, H. && (6) \end{aligned}$$

In this setting, the term $(\sum_{r=1}^{R} a_{jr} z_{rh})$ contains those factors that for each household are common across goods but do not depend on total expenditure, that is those forces other than income that affect households in their own way. Some studies have investigated the restrictions that consumer theory imposes not only on the shape of Engel curves but also on the structure of errors (Lewbel, 2001; Blundell et al., 1998, 2003, 2007). In particular, although the class of compatible demand models is very flexible in terms of functional forms, still for any curve $g_r(x_h)$ we cannot interpret the error terms z_{rh} as random preference parameters or individual location shifts. Therefore, we must allow for correlation across households in z_{rh}. In sections 5 and 6, we make specific hypotheses about the statistical properties of these error terms, proposing some intuitive economic interpretations, but remaining agnostic about its possible interpretation on the basis of consumer theory.

3 Building the dataset

In order to estimate model (4), we need data on how a sample of families has allocated the budget across different categories of expenditures. This dataset has to fulfil some specific requirements which permit us to apply factor and independent components analysis (the statistical reason behind these requirements will be apparent in the next section). First of all, we need to deal with a large panel: both dimensions — in our case the number of households and the number of categories of expenditure — have to be high. Moreover, the panel has to be perfectly balanced, that is we want to know how each household allocates its budget for each selected category of expenditure.

These two requirements are not easy to be simultaneously fulfilled in standard expenditure national surveys because usually we have complete information as to how a large sample of households allocated their expenditures towards a limited number of categories of expenditure. In order to get a large number of expenditures, one option could be to look at numerous disaggregated categories; expenditure surveys often keep track of these values. However, these values are not as reliable as the macro–categories and that there is the problem of zero expenditures, since for each micro–category there is always a number of households whose corresponding expenditure is zero or missing.

Considering that expenditure surveys are regularly repeated on an annual basis, another option is to pool together data collected in different years. In this manner we can keep using macro–categories, but at the same time we can considerably increase the number of expenditure categories, since we have a set of macro-categories for each year. This is the route we take. There are, however, some issues related to this approach. First, when considering expenditure data over different years, we have to control for the fact that prices for each category of expenditure have changed. We tackle this problem by converting nominal values to real values of expenditures using category–specific price indices. Second, we cannot keep track of single households. We address this problem by examining average allocations among groups of income–homogeneous households. For each year, we divide the data in 100 intervals based on the average percentiles of the distribution of total expenditure. By averaging expenditures within each interval we obtain for each year a class of $H = 100$ representative households. In this way, for each representative household, we are able to observe its expenditure allocations over several years. Thus, for example, corresponding to the household representative of the h^{th} interval we can observe its expenditure allocation towards the category of expenditure g at time t, $t + 1$, etc. This procedure relies on the assumption that individual demands are exactly aggregable within each interval and that total expenditure is constant over time.

We use data from the U.K. Family Expenditure Survey (FES) 1977–2001 jointly with the Expenditure and Food Survey (EFS) 2002–2006. We have data about household expenditures on various categories of goods and services. Each year approximately 7000 households were randomly selected, and each of them recorded expenditures for two weeks. We are able to recover information about total expenditures and expenditures on fourteen aggregated categories: (1) housing (net); (2) fuel, light, and power; (3) food; (4) alcoholic drinks; (5) tobacco; (6) clothing and footwear; (7) household goods; (8) household services; (9) personal goods and services; (10) motoring, fares and other travel; (11) leisure goods; (13) leisure services; and (14) miscellaneous and other goods. The 14 categories add up to total expenditure. We omit from our analysis the last category of expenditure and we restrict therefore to $G = 13$ categories. A description of the disaggregated categories of expenditure included in each of the 13 classes is available in the complementary appendix. In order to have samples of households which are demographically homogeneous, we control for the number of members of each household and we consider four different possibilities: 1 member, 2 members, 2 or 3 members, and 2 to 4 members.[3] The sizes of these samples range from 1700 (1 member group in 1997) to 4844 (2–4 members group in 2002). Finally, we pool together budget shares over different years, choosing three different waves of $T = 10$ years each: 1977–1986, 1987–1996, and 1997–2006. Thus, we are able to get $J = GT = 130$ budget shares for each wave considered. The procedure to build the data set, which is similar to the one adopted by Kneip (1994), is described in detail in table 1.

The approach followed to build the dataset allows us to estimate model (4), which in turn is justified if: *i*) the distribution of real total expenditure is stable across the considered years; and *ii*) the structure and number of latent Engel curves is not changing over time. In the next section we address both these stability issues.

4 Preliminary data analysis

In this section, we analyze data separately for each year considered. We perform three analysis: first we study the distribution of total expenditure, second we determine the number of common factors in each year, and third we compare the estimated basic Engel curves or common factors, across years.

The distribution of total expenditure. According to our approach each representative household refers to an interval of total expenditure whose boundaries are obtained by taking the average of percentiles over time. Thus we have to study the distribution of total expenditure. In figure 1, we show the Box–plots for this distribution in each year from 1997 to 2006 and for the four samples considered. It can be appreciated that the median, the 25th and 75th percentile, and the maximum and minimum values are fairly stable during the whole period.

Number of factors. For each year t considered, the deflated model (5), is

$$w_{gh\,t} = \mathbf{a}'_{g\,t}\mathbf{f}_{h\,t} + \ e_{gh\,t}, \quad g = 1, \dots, G; \ \ h = 1, \dots, H, \ \ t = 1, \dots, T, \tag{7}$$

where $\mathbf{a}_{g\,t}$ and $\mathbf{f}_{h\,t}$ are R_t–dimensional vectors of loadings and latent factors respectively and $e_{gh\,t}$, is a G dimensional vector of mean zero errors that are assumed to be independent of total expenditure $x_{h\,t}$. For any t, we then define the $G \times H$ budget shares matrix $\mathbf{w}_t$, the $G \times R_t$ loadings matrix $\mathbf{A}_t$, the $R_t \times H$ factor matrix $\mathbf{f}_t$, and the $G \times H$ errors matrix $\mathbf{e}_t$.

In order to determine R_t we cannot make use of the approximate principal component analysis

[3] Controlling for the age of the head of household would reduce too much the number of available observations.

outlined in the next section as such estimators deliver consistent estimates only if both the sample size and the cross–sectional dimensions are large. This is not the case here, since for each year we have observations only for $G = 13$ categories of expenditure. Moreover, tests for the number of factors in classical factor analysis cannot be used as the error terms are likely to be correlated.

Following Lewbel (1991), in order to estimate R_t, we can exploit the fact that the factors are all functions of total expenditure while errors are independent of the factors and therefore are independent of total expenditure. For each t and h, let $\mathbf{q}(x_{h\,t}) = (q_1(x_{h\,t}), \dots, q_G(x_{h\,t}))'$ be a G–dimensional vector of functions of total expenditure having finite means and denote by $\mathbf{Q}(\mathbf{x}_t)$ the corresponding $G \times H$ matrix containing the G functions of total expenditure for every household. Then, from (7) we define the $G \times G$ matrix

$$\mathbf{w}_t\mathbf{Q}(\mathbf{x}_t)' = \mathbf{A}_t\mathbf{f}_t\mathbf{Q}(\mathbf{x}_t)' + \mathbf{e}_t\mathbf{Q}(\mathbf{x}_t)', \quad t = 1, \dots, T. \tag{8}$$

Then, since by assumption $\mathrm{E}[\mathbf{e}_t\mathbf{Q}(\mathbf{x}_t)'] = \mathbf{0}$ at any t, we have that $\mathbf{Y}_t = \mathrm{E}[\mathbf{w}_t\mathbf{Q}(\mathbf{x}_t)'] = \mathbf{A}_t\mathrm{E}[\mathbf{f}_t\mathbf{Q}(\mathbf{x}_t)']$ has rank R_t, unless by coincidence some component of the factors is orthogonal to all the elements of $\mathbf{Q}(\mathbf{x}_t)$, in which case we would have a smaller rank.

Thus, in order to determine R_t we can test for the rank of $\mathbf{Y}_t$ at each point in time. Lewbel (1991) proposes a way to test for the rank of $\mathbf{Y}_t$ based on the LDU decomposition of its sample counterpart $\widehat{\mathbf{Y}}_t$ which has generic (i, j)–th entry $\widehat{Y}_{ij\,t} = \frac{1}{H}\sum_{h=1}^{H} w_{ih\,t}q_j(x_{h\,t})$. The test is for the null–hypothesis of a rank equal to R_t against the alternative of a rank greater than R_t and has an asymptotic distribution which is $\chi^2_{(G-R_t)}$. We refer to the original paper for details on how to build the test. Results for all datasets considered are in table 2.[4] Results for this test are in table 2 and denote an almost constant number of common factors across time which is between two and three on average in agreement with the benchmark case considered below where we set $R_t = 3$. A similar result is obtained also by Lewbel (1991) for the period 1970–1984. The conclusion is that the number of basic Engel curves has remained almost constant during the period considered.

Factors' space. In order to justify the pooling of different waves of budget shares, we have to show that factors did not change over time. Here we have two difficulties when dealing with small cross–sections. First, we cannot estimate factors consistently unless we assume a diagonal covariance matrix of the residuals. Second, even when an estimate of the factors is available, single factors are not identified. Concerning estimation we build a covariance matrix which has non–zero elements only on the T diagonal blocks of size $G \times G$, and we denote it by $\mathbf{\Sigma}^{\mathbf{w}}_{b\,t}$. This is

[4] The 13 functions $q_j(x_h)$ considered are: x_h^a, $\log x_h^a$, with $a = \pm 1, \pm 2, \pm 3$, and $x_h \log x_h$.

like imposing zero covariance across budget shares of different years. In this way, we can estimate R_t factors for each year by estimating the largest TR_t principal components of $\Sigma^{\mathbf{w}}_{b\,t}$.

In particular, for $j = 1, \dots, R_t$, the first R_t factors of block t are obtained by projecting $\mathbf{w}_t$ onto the space spanned by the $(j-1)T + t$ largest eigenvectors of $\Sigma^{\mathbf{w}}_{b\,t}$. We denote such estimated factors as $\widehat{\mathbf{f}}_{b\,t}$ which is an $R_t \times H$ matrix and for any t we compare them with those estimated on the pooled dataset and denoted as $\widehat{\mathbf{f}}$ which is an $R \times H$. Since factors are not identified we can only compare the space they span and this is done by means of the following statistics (see e.g. Doz et al., 2012):

$$\tau_t = \frac{\text{trace}\left(\widehat{\mathbf{f}}\ \widehat{\mathbf{f}}'_{b\,t}\left(\widehat{\mathbf{f}}_{b\,t}\,\widehat{\mathbf{f}}'_{b\,t}\right)^{-1}\widehat{\mathbf{f}}_{b\,t}\,\widehat{\mathbf{f}}'\right)}{\text{trace}\left(\widehat{\mathbf{f}}\,\widehat{\mathbf{f}}'\right)},$$

which is a multivariate version of the R^2 coefficient of the regression of $\widehat{\mathbf{f}}$ on $\widehat{\mathbf{f}}_{b\,t}$. We compute this measure for any block t and since in the pooled case the number of factors R does not depend on t we have to fix $R = R_t$ and we compute the measure for different values of R. Results for $R = 1, \dots, 4$ are in table 3. In most of the cases considered, the value of τ_t is about 0.90 and it is even higher for the benchmark case. This result indicates that the estimated factors on single blocks (years) span always the same space as those estimated on the pooled data.

5 An approximate factor model for budget shares

As suggested by Bai and Ng (2002), we can consider equation (4) as a factor model with R factors common to the $J = GT = 130$ budget shares, where $R < J$. Thus, for every household h we can write the budget share for expenditure category j as in (5):

$$w_{jh} = \mathbf{a}'_j\mathbf{f}_h + \ e_{jh}, \quad j = 1, \dots, J; \;\; h = 1, \dots, H, \tag{9}$$

where $\mathbf{a}_j$ and $\mathbf{f}_h$ are R–dimensional vectors of loadings and latent factors respectively. In matrix notation

$$\mathbf{w} = \mathbf{A}\mathbf{f} + \ \mathbf{e}, \tag{10}$$

where $\mathbf{w}$ and $\mathbf{e}$ are $J \times H$, $\mathbf{A}$ is $J \times R$, and $\mathbf{f}$ is $R \times H$. We call the term $\mathbf{Af}$ the common component and the term $\mathbf{e}$ the idiosyncratic component orthogonal to the factors. While an *exact* factor model would require that idiosyncratic components are uncorrelated across expenditure categories, this is here an unreasonable restriction. We cannot exclude correlated idiosyncratic components because budget shares as this a direct consequence of the adding up condition in a system of Engel curves

(Lewbel, 1991). Although we have omitted in our study the category "miscellaneous goods", so that budget shares do not exactly add up to T, still budget shares for this category have a negligible contribute to the total budget. Therefore, even if adding up is not exactly fulfilled, we cannot exclude the existence of non–zero covariances in the idiosyncratic components. More in general, idiosyncratic components are likely to be correlated because they capture good–specific influences other than income and nothing excludes dependencies among them.

On the other hand, in *approximate* factor models a large J allows for mildly correlated idiosyncratic terms. In fact, a large cross–section of budget shares is what allows us to choose a different modelling and estimation strategy with respect to Lewbel (1991). Namely, we can apply the theory by Bai and Ng (2002) in this paper. The necessity of having a large number of items is the practical reason for pooling expenditures of different years together when building the dataset as described in section 3 while the empirical justification for this approach is provided in section 4. The complete details and assumptions for the approximate factor model are in Bai and Ng (2002) and we recall here just the three main assumptions:

1. factors: $\lim_{H\to\infty} \frac{1}{H}\sum_{h=1}^{H} \mathbf{f}_h\mathbf{f}_h' = \mathbf{\Sigma}^{\mathbf{f}}$, for some positive definite and diagonal $R\times R$ matrix $\mathbf{\Sigma}^{\mathbf{f}}$;
2. loadings: $\lim_{J\to\infty} ||\mathbf{A}'\mathbf{A}/J - \mathbf{D}|| = 0$, for some positive definite $R\times R$ matrix $\mathbf{D}$[5];
3. idiosyncratic components: define $\mathbf{\Sigma}^{\mathbf{e}} = \mathrm{E}[\mathbf{e}_h\mathbf{e}_h']$ then there exists $M > 0$ s.t. $\sum_{k=1}^{J} |(\mathbf{\Sigma}^{\mathbf{e}})_{jk}| \leq M$ for any $j = 1, \dots, J$.

Assumption 1 implies the existence of the covariance matrix of the factors which being diagonal implies that the factors are orthogonal. Assumption 2 is sufficient for identification of the loadings and implies that, when J goes to infinity, $\mathbf{A}'\mathbf{A}$ is $O(J)$. Assumption 3 defines an approximate factor model by allowing for some correlation across goods in the idiosyncratic components, this is equivalent to require the largest eigenvalue of $\mathbf{\Sigma}^{\mathbf{e}}$ to be bounded as J goes to infinity (see also Chamberlain and Rothschild, 1983).

The rank of the considered system of budget shares, is therefore the smallest integer R such that equation (9) holds. While Lewbel (1991) proposes a test based on LDU decomposition to determine R, both Kneip (1994) and Donald (1997) propose non–parametric estimation methods. We instead adopt here the estimation method proposed by Bai and Ng (2002), based on approximate principal component analysis. This approach provides a consistent estimate of R and the

[5] We use the Froboenius norm for a matrix, i.e. $||\mathbf{B}|| = \sqrt{\mathrm{tr}(\mathbf{B}\mathbf{B}')}$.

space spanned by the factors when both H and J go to infinity. In the rest of this section we first briefly review how to estimate factors via principal components and how to determine R. We then provide an identification strategy based on statistical independence. Finally, from (5) and (6), we see that the elements of the R-dimensional vector of basic Engel curves $\mathbf{g}(x_h)$ may be recovered by regressing each identified factors on total expenditure.

Estimation. First let us assume that R is known, then the estimated factors and loadings are obtained by solving

$$(\widehat{\mathbf{f}}, \widehat{\mathbf{A}}) = \arg\min_{(\mathbf{f},\mathbf{A})} V(R, \mathbf{A}, \mathbf{f}) = \arg\min_{(\mathbf{f},\mathbf{A})} \frac{1}{JH} \sum_{j=1}^{J} \sum_{h=1}^{H} (w_{jh} - \mathbf{a}_j' \mathbf{f}_h)^2, \tag{11}$$

subject to an additional identification condition which consistently with assumption 2, we require to be $\widehat{\mathbf{A}}'\widehat{\mathbf{A}}/J = \mathbf{I}_R$, where $\mathbf{I}_R$ is the R-dimensional identity matrix. With this choice, the columns of $\widehat{\mathbf{A}}$ are given by $\sqrt{J}$-times the eigenvectors corresponding to the R largest eigenvalues of the sample covariance matrix of the observed budget shares $\frac{1}{H}\sum_{h=1}^{H} \mathbf{w}_h \mathbf{w}_h'$, where $\mathbf{w}_h$ is the J-dimensional vector of budget shares of household h. In the limit $J, H \to \infty$ the estimated loadings $\widehat{\mathbf{A}}$ are a consistent estimate of $\mathbf{A}$ and the factors can be consistently estimated as the R largest principal components: $\widehat{\mathbf{f}} = \widehat{\mathbf{A}}'\mathbf{w}/J$ (see Theorem 1 in Bai and Ng, 2002, for a proof).

Following Bai and Ng (2002), we can use the above estimation method to estimate the number of factors R. This can be done by estimating the factors and their loadings for different values k of the number of factors and by solving each time (11). Define $\widehat{\mathbf{A}}^k$ and $\widehat{\mathbf{f}}^k$ as the approximate principal components estimates of loadings and factors when assuming the existence of k common factors. The estimated number of factors is the value of k that minimizes this function, conveniently penalized with a penalty function $p(k, J, H)$ that depends both on J and on H. We thus look for minima of the ICs criteria proposed by Bai and Ng (2002), i.e.

$$\widehat{R} = \arg\min_{1 \le k \le k_{\max}} \log V(k, \widehat{\mathbf{A}}^k, \widehat{\mathbf{f}}^k) + p(k, J, H) \tag{12}$$

where

$$\begin{aligned} p(k, J, H) &= k\left(\frac{J+H}{JH}\right)\log\left(\frac{JH}{J+H}\right) \\ &\text{or} \\ p(k, J, H) &= k\left(\frac{J+H}{JH}\right)\log\left(\min\left(\sqrt{J}, \sqrt{H}\right)\right)^2. \end{aligned} \tag{13}$$

Provided that we have a consistent estimate of the factors and their loadings, Bai and Ng (2002) prove consistency of $\widehat{R}$ as $J, H \to \infty$. In the following sections we also apply three other methods:

i) a refinement of the above information criteria proposed by Alessi et al. (2010) where a fine–tuning parameter in the penalty function is introduced; *ii*) test by Onatski (2010) which is instead based on the asymptotic distribution of the eigenvalues of the sample covariance matrix; *iii*) the test based on conditional correlations and presented in section 4 but applied to the pooled dataset.

Identification. Factor models have an indeterminacy which they cannot solve: both the estimated loading matrix $\widehat{\mathbf{A}}$ and factors $\widehat{\mathbf{f}}$ are asymptotically consistent estimates of the true ones only up to an orthogonal transformation. We have, therefore, an identification problem which makes difficult the economic interpretation of the estimated factors. In order to identify the model, we use independent component analysis (ICA) which requires two further assumptions on the R latent factors:

4. the components of the factor vector $\mathbf{f}_h$ are mutually independent, i.e. the joint probability density of the factors is given by

$$\mathcal{D}(\mathbf{f}_h) = \prod_{r=1}^{R} d_r(f_{rh}), \quad h = 1\ldots, H,$$

where d_r is the marginal probability density of the r-th factor;

5. the marginal densities d_r are non–Gaussian, for all $i = 1, \ldots, R$, with the exception of at most one.

Assumption 4 is justified on the basis of the fact that the latent factors represent the basic latent Engel curves generating the observed system of Engel curves. These basic functions, in turn, have characteristics which reflect fundamental aspects of human behaviours driving consumption decisions. As argued by Witt (2001), consumption decisions are ultimately driven by basic needs and acquired wants. Therefore, assuming that latent factors are independent amounts to claim that the set of needs and wants associated with each factor is of fundamental different nature, i.e. generates an independent pattern, from the set of needs and wants associated with the other factors. For example, if a factor reflects a pattern associated with necessities and another factor reflects a pattern associated with luxuries, these two factors can be seen as statistical independent, because necessities mainly reflect physiological needs, while luxuries reflect culturally acquired wants such as social recognition and status. The drivers underlying consumption decisions about necessities and luxuries react in an independent way to changes in income: for example, physiological needs tend rapidly to satiate, as income gives the possibility to satisfy these needs, whereas acquired wants such as social recognition and status may be even increasingly reinforced, as income increases.

Nevertheless, it has to be stressed that while basic Engel curves reflect independent motives for consumption, the observed Engel curves can be seen as a mixture of these needs and thus their joint distribution may have a non–trivial dependence structure.

Assumption 5 is justified by testing for normality in the data and also by noticing that often data on consumption expenditures are non-Gaussian (see e.g. Fagiolo et al., 2010) and, moreover, being budget shares defined on the unit interval, they must have a distribution with bounded support (e.g. a beta distribution) hence not a Gaussian distribution. As a consequence also the joint distribution of the factors is non–Gaussian.

ICA can been seen as an extension or a strengthening of principal component analysis (PCA) (see Comon, 1994; Hyvärinen et al., 2001; Bonhomme and Robin, 2009). Indeed, while PCA gives a transformation of the original space such that the computed latent factors are linearly uncorrelated, ICA goes further by attempting to minimize all statistical dependencies between the resulting components. One can show that *if* there exists a representation with non-Gaussian, statistically independent components, then the representation is essentially unique (up to a permutation, a sign, and a scaling factor) (Comon, 1994). There exist a number of computationally efficient algorithms for consistent estimation (Hyvärinen et al., 2001). This identification method is particularly appealing since it is purely data–driven and not based on economic assumptions which in turn would require micro–funded models of consumption behavior.

The most popular ICA algorithms are: Joint Approximate Diagonalization of Eigen-matrices (JADE by Cardoso and Souloumiac, 1993), Fast Fixed-Point Algorithm (FastICA by Hyvärinen and Oja, 2000). Both methods are based on two steps: *i*) a whitening step achieved by PCA, in which the data are transformed so that the covariance matrix is diagonal and has reduced rank, i.e. we get rid of the idiosyncratic component; *ii*) a source separation step in which the orthogonal transformation necessary for achieving identification is determined.

When data usually tend to exhibit fat-tailed distributions and poor serial correlation (in our framework we have no correlation at all across households), JADE and FastICA which are based on non-Gaussianity of the data, hence on higher order moments, are the most used algorithms.[6] We present here results obtained with JADE, the results obtained with FastICA being similar.

Once estimation of the common component is accomplished via approximate PCA, we are left

[6] Another algorithm is Second-Order Blind Identification (SOBI Belouchrani et al., 1997), which, although usually applied in time-series analysis, could be extended to cross-sectional data with correlations among observations. However, this is not the case for us, as we assume no correlations across households.

with a first estimate of the factors $\widehat{\mathbf{f}}_h$ for any household h. JADE looks for an orthogonal $J \times R$ matrix $\widehat{\mathbf{U}}$ such that the identified factors $\widetilde{\mathbf{f}}_h = \widehat{\mathbf{U}}'\widehat{\mathbf{f}}_h$ are maximally non-Gaussian distributed. A set of random vectors is mutually independent if all the cross-cumulants (i.e. the coefficients of the Taylor series expansion of the log of the moment generating function) of order higher than two are equal to zero. In particular, Cardoso and Souloumiac (1993) prove that the factors $\widetilde{\mathbf{f}}_h$ are maximally independent if their associated fourth-order cumulant tensor which is a $R \times R$ matrix is maximally diagonal.[7] JADE is a very efficient algorithm in low dimensional problems as the one treated here (we have few factors), while a higher computational cost is required when the dimension increases.

Once we apply JADE the estimated and identified factors, $\widetilde{\mathbf{f}}_h$, are identified up to a permutation, a sign, and a scaling factor. The order of the factor is irrelevant for our purposes. Moreover, given that independent components are nothing else but weighted averages of the data, the sign is chosen to be consistent with the average of budget shares across goods. Finally, the scale is determined in such a way that the identified loadings $\widetilde{\mathbf{A}}$ satisfy $\widetilde{\mathbf{A}}'\widetilde{\mathbf{A}}/J = \mathbf{I}_R$.

6 Estimation of the basic Engel curves

As shown in section 2, for each latent factor we consider the following model

$$f_{rh} = g_r(x_h) + \ z_{rh}, \quad r = 1, \dots, R, \tag{14}$$

where we introduced an error term in the specification of the latent factors. We make the following assumption on the vector $\mathbf{z}_r = (z_{r1} \dots z_{rH})'$:

6. for any $h, \ell = 1 \dots H$, $\mathrm{E}[z_{hr}] = 0$, $\mathrm{E}[z_{hr}^2] = \sigma_h^2$, and there exist a constant $\zeta > 1$ and a function $\rho(\cdot)$ such that $\mathrm{Corr}(z_{rh}, z_{r\ell}) = \rho(|x_h - x_\ell|) \sim |x_h - x_\ell|^{-\zeta}$.

Assumption 6 implies that the correlation between errors of two households depends on the difference between their total expenditure, and such dependence decreases at a rate given by ζ (see e.g. Härdle, 1990). The z_{hr} terms captures measurement errors and expenditure influences that are both other than income and common across goods. Such influences are likely to be correlated across households (the other–than income influence affecting household h can be correlated with

[7] While the cumulant depends on four indexes the cumulant tensor depends on two indexes, the other two being canceled by means of an additional arbitrary matrix. We thus have to consider several cumulant matrices which have to be jointly diagonalized. See the appendix for a short description of the JADE algorithm.

the other–than income influence affecting household ℓ), but we assume here that this dependence wanes out as the income difference between households becomes big enough. This is consistent with the intuition that other-than income influences such as fashions or technical change affect “distant” income classes of households in their particular manner. This assumption is necessary in the non–parametric setting but it can be relaxed in the parametric case.

The aim of this section is to provide consistent estimators of the basic Engel curves $g_r(x_h)$. In what follows we propose a non–parametric and a parametric estimator of these curves. While the former is appealing since it is purely data driven, the latter allows us to relate our results with the existing literature on functional forms of Engel curves (see e.g. Lewbel, 1991; Banks et al., 1997).

Proposition 1. *The estimator for the basic Engel curve $g_r(x_h)$ is defined as $\widetilde{\gamma}_r^*(x_h)$, such that*

$$\widetilde{\gamma}_r^*(x_h) = \arg\max_{\gamma_r} \sum_{k=1}^{H} \left[\widetilde{f}_{rk} - \gamma_r - \delta_r(x_k - x_h)\right]^2 \mathrm{K}_{b_H}(x_k - x_h), \quad r = 1, \ldots, R, \tag{15}$$

where $\mathrm{K}_{b_H}(\cdot)$ is a suitable kernel function depending on a bandwidth b_H (see assumption K in the appendix). Then, under assumptions 1–6,

$$p\text{-}\lim_{J,H\to\infty} |\widetilde{\gamma}_r^*(x_h) - g_r(x_h)|^2 = 0,$$

with a rate of convergence given by $\min\left(J^{-1}, H^{-1}b_H^{-1}, b_H H^{-1}\right)$.

Proof: see the appendix.

Few remarks are necessary. First, since f_{rh} are unobserved and must be replaced by their estimates $\widetilde{f}_{rh}$, we have to use lemma 1 in the appendix and consistency is achieved provided that both H and J tend to infinity. Second, the proposed estimator is a local linear estimator as defined for example in Fan and Gijbels (1992) and Fan (1993). An alternative estimator is represented by the local constant fit defined as,

$$\widetilde{\gamma}_r^*(x_h) = \arg\max_{\gamma_r} \sum_{k=1}^{H} \left[\widetilde{f}_{rk} - \gamma_r\right]^2 w_k(b_H), \quad r = 1, \ldots, R. \tag{16}$$

From (16), we can have either the Nadaraya–Watson estimator when $w_k(b_H) = \mathrm{K}_{b_H}(x_k - x_h)$ or the Gasser–Müller estimator when $w_k(b_H) = \int \mathrm{K}_{b_H}(u - x_h)\mathrm{d}u$ (see Watson, 1964; Gasser and Müller, 1984, respectively). Both (15) and (16) would satisfy Proposition 1. However, it can be proved that the local linear estimator (15) has a smaller finite sample bias, is asymptotically efficient, and has a better behavior at the extremes of the sample (see Fan and Gijbels, 2003, for a comparison). Moroever, by solving the maximization in (15), we obtain also a local estimate of the slope $\widetilde{\delta}_r^*(x_h)$ which is an estimate of the first derivative of the basic Engle curves. Consistency of the latter in our framework is proved exactly in the same way as in Proposition 1. Finally, the presence of correlation in the error terms does not affect the rate of consistency as long as we have weak dependence as given by assumption 6.

The choice of the bandwidth can be based on different methods. In our estimations below, we choose the bandwidth on the basis of the minimization of a polynomial approximation of the mean integrated square error (of $\widetilde{\gamma}_r^*(x_h)$), following the approach proposed by Fan and Gijbels (2003, Section 4.2). The presence of correlation in the errors has also to be taken into account when selecting the bandwidth as proposed in Altman (1990).

In order to compare our results with the literature (Lewbel, 1991; Banks et al., 1997), we also investigate which functional form of total expenditure better fits each identified factor. Thus instead of (15) we can think of a parametric model for the basic Engel curves:

$$g_r(x_h) = \alpha_r + \beta_r m(x_h), \quad r = 1, \dots, R; \;\; h = 1, \dots, H. \tag{17}$$

We estimate the following functions $m(x_h)$ of total expenditure: x_h, x_h^2, x_h^{-1}, x_h^{-2}, $\log x_h$, $(\log x_h)^2$, $x_h \log x_h$. These are are the functional forms also considered by Lewbel (1991) and Donald (1997). By substituting (17) into (14) we have

$$f_{rh} = \alpha_r + \beta_r m(x_h) + \; z_{rh}, \quad r = 1, \dots, R; \;\; h = 1, \dots, H. \tag{18}$$

The unknown parameters can be estimated by ordinary least squares with the caveat that, since z_{rh} are non–Gaussian by assumption 5, robust standard errors must be computed. If the factors f_{rh} were observed, consistency of the estimated parameters would follow from Quasi Maximum Likelihood theory. However, since f_{rh} are unobserved and must be replaced by their estimates $\widetilde{f}_{rh}$, we have to use lemma 1 and consistency is achieved provided that both H and J tend to infinity. We have the following result.

Proposition 2. *Define the matrix of explanatory variables and the vector of unknown parameters*

$$\mathcal{X} = (\mathbf{1}_H, m(\mathbf{x})), \qquad \boldsymbol{\theta}_r = \begin{pmatrix} \alpha_r \\ \beta_r \end{pmatrix}, \quad r = 1, \dots, R,$$

where $\mathbf{1}_H$ is an H-dimensional column vector of ones and $\mathbf{x} = (x_1 \dots x_H)'$. The estimated vector parameters for the r-th basic Engel curve is given by

$$\widetilde{\boldsymbol{\theta}}_r^* = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\widetilde{\mathbf{f}}_r,$$

such that, under assumptions 1–5,

$$p\text{-}\lim_{J,H\to\infty} |\widetilde{\boldsymbol{\theta}}_r^* - \boldsymbol{\theta}_r| = 0, \quad r = 1, \dots, R,$$

with a rate of convergence given by $\min\left(J^{-1}, H^{-1}\right)$.

Proof: see the appendix.

Notice that assumption 6 of weak dependence is not necessary for the parametric estimates. Both in the (15) and in the parametric (17) case consistency of the estimated conditional means is not affected by the correlation across errors, however their standard errors are affected. In what follows we compute standard errors from a distribution of 1000 fits obtained by estimating and identifying the factors on boostrapped samples of the observed budget shares. In this way we account also for the error made in estimating the factors and moreover we do not need to compute the analytical form of the errors of the estimated parameters.

7 Results

Average budget shares. In table 4, we report the average (across households) budget shares for all 13 considered expenditure categories and the four different demographic groups. The majority of the budget (about 20%) is spent for food and housing followed by motoring and leisure services (about 10%). A smaller fraction of budget is allocated to all other goods with percentages that do not reach the 10%. We notice that the average food budget share remains constant across demographic groups, while there is some heterogeneity in housing and motoring expenditure: single member households allocate a higher proportion of household budget towards housing, and a lower proportion of budget toward motoring in comparison in comparison with the other demographic groups. Also poorer households spend 10% more for food than rich ones, while motoring expenditure is 5% higher for richer households. These across–income differences is what we want to model and estimate in this paper by means of latent Engel curves. Indeed, already from such descriptive analysis we can tentatively classify goods according to their budget shares into three broad classes: necessities (budget shares decreasing with total expenditure), luxuries (budget shares increasing with total expenditure), and goods for which the budget share is constant with respect to total expenditure.

Number of factors. Table 5 displays the estimates of the number of factors. Beside the time window 1997-2006, we report here, for the sake of comparison, also results for the time windows 1977-1986 and 1987-1996. We find between 2 and 4 common factors, the average of the criteria being always about 3. In the following analysis the main role is played by the first two largest factors, while a third plays a minor although theoretical important role. Adding more factors does not change the interpretation and therefore we present results for $R = 3$. In the last 3 rows of table 5 we also show the proportion of variance explained by each factor. The first factor explains, for most of the samples considered, between 50% and 70% of total variance, being clearly the most important, while the second explains about 10%.

Factor models are identified under a specific condition on diverging eigenvalues of the covariance matrix of the data (see assumption 3). This is precisely the assumption tested by the Bai and Ng (2002) and Alessi et al. (2010) criteria. It has to be noticed that often in the literature it is found that the Bai and Ng (2002) criterion tends to overestimate the number of factors, while the one by Alessi et al. (2010) being based on the tuning of the penalty tends to perform better. We find evidence of one or two additional factors less important, but still common and explaining a much lower proportion of variance, in fact lower than 5%. We must stress the fact that not recognizing the existence of such factors would imply the existence of common features in the idiosyncratic components. Indeed, in order to be truly common the factors do not have to be necessarily large (a relative concept) in terms of explained variance, but they have to be pervasive, a well defined feature that can be measured by studying the asymptotic behaviour of eigenvalues. This is exactly what the employed criteria do. Similar results are obtained by employing the criterion by Onatski (2010). Finally, a higher number of factors is obtained when using the test based on canonical correlations and described in section 4.[8]

Factor identification and estimates of basic Engel curves. Hereafter, we present results only for the last 10 years window considered, i.e. from 1997 to 2006 and for households with 2 to 4 members.[9] The identification of the factors is based on the independent component analysis, as explained in section 5. This method can be applied only if the underlying independent components, and, consequently, the estimated (non–identified) factors are non-Gaussian. Figure 2 shows the quantiles of first two estimated factors *vs.* Gaussian quantiles: a non–linear relation clearly appears. This suggests that at least two out of three estimated factors do not follow a Gaussian distribution and this is enough to allow for identification via JADE. We also test directly for Gaussianity. The Shapiro-Wilk test rejects the hypothesis of Gaussianity at the 5% level of significance for both the three factors estimated via PCA and for the identified factors.

As a preliminary analysis of the meaning of the identified factors, we report in table 6 the estimated factor loadings for each category of budget shares averaged over 10 years. These are

[8] In this case since we deal with the large dataset we cannot compute the LDU decomposition and the related test due to computational reasons.

[9] Additional results regarding other samples are available in the complementary appendix.

a measure of the correlation between the observed budget shares and the identified factor. As explained above, the scale of the loadings vector is fixed according to the normalization $\widetilde{\mathbf{A}}'\widetilde{\mathbf{A}}/J = \mathbf{I}_R$. We find that the first factor is highly correlated with food and fuel, light and power budget shares. This again suggests that the first factor captures consumption patterns typically associated with the Engel's law: as total expenditure rises, budget shares decrease, the downward trend being more dramatic for the lowest levels of income. On the other hand the second factor is mostly correlated with luxuries as motoring and leisure services, while the third displays the highest correlation with food and housing expenditures.

We then consider regressions in order to estimate the basic Engel curves. Figure 3 (a-c-e) displays the three factors $\widetilde{f}_{rh}$ (represented by circles) as functions of total expenditure together with their estimated fits $\widetilde{\gamma}_r^*(x_h)$, as obtained by means of the local linear kernel regression, as described in section 6. Estimates are reported together with their 68% and 90% confidence intervals based on the standard errors of a distribution of 1000 fits obtained by estimating and identifying the factors on boostrapped samples of the observed budget shares. In this way we account for the error made in the estimation both of the factor and of the regression. The first function $\widetilde{\gamma}_1^*(x_h)$ decreases for small values of total expenditure and then remains stable. This pattern is very similar to the pattern of food and fuel budget shares, as evidenced from figure 4 (a-b). The second function, $\widetilde{\gamma}_2^*(x_h)$ is increasing with total expenditure, apart from the first portion of total expenditure. It is associated with categories of expenditure which are more likely to include luxuries as clothing and footwear, motoring, and leisure services. Indeed, from figures 4 (c-d) we see that the second factor displays a pattern similar to leisure service and motoring budget shares. Finally, the third function, $\widetilde{\gamma}_3^*(x_h)$, is slightly increasing in the first quarter of total expenditure and then slightly decreasing, remaining on average approximately constant. This pattern is similar to the one displayed by housing and alcoholic drinks (see figure 4 (e-f)).

As explained in section 6, we also investigate which functional form of total expenditure better fits each identified factor. Following Lewbel (1991) and Donald (1997), we consider the following functions of total expenditure: x_h, x_h^2, x_h^{-1}, x_h^{-2}, $\log x_h$, $\log^2 x_h$, $x_h \log x_h$. In this way, we can compare our results with the literature. In table 7, we show the adjusted R^2 coefficient for the different functional forms. The first Engel curve captures most of the variation in budget shares of poor households while the second captures most of the variation in budget shares of rich households. Notice, however, that, the R^2 coefficient for the third factor, is quite small for most of the functional forms considered, so that a constant relation constitutes a good approximation as also shown by figure 3 (e).

Moreover, by inspecting the estimated coefficients in table 8, we see that first Engel curve is best represented by the an inverse or a logarithmic form with negative slope, i.e. by a monotonically decreasing curve. In particular, the logarithmic functional form is incorporated in the Working–Leser model. The best representation of the second Engel curve is given by a quadratic form either x_h^2 or $x_h \log x_h$, thus increasing for large values of total expenditure. Finally, we notice that the first factor is the most important when considering only poor households, while the second prevails when considering households with medium–high levels of income.

Summing up, among all the possible parametric form considered, our findings are for example consistent with the following parametric specification:

$$w_{jh} = c_{1j} + c_{2j} \log x_h + c_{3j} x_h \log x_h + \eta_{jh}, \quad j = 1, \dots, J, \quad h = 1, \dots, H, \tag{19}$$

where c_{rj} are combinations of the loadings a_{jr} and the coefficients α_r and β_r and η_{jh} contains both the idiosyncratic components e_{jh} and the terms $a_{jr} z_{rh}$ representing common features of budget shares not due to total expenditure. The constant term represents the third factor. The functional form (19) is consistent with the one proposed by Lewbel (1997):

$$w_{jh} = c_{1j} + c_{2j} \log x_h + c_{3j} \psi(x_h) + \eta_{jh}, \quad j = 1, \dots, J, \quad h = 1, \dots, H, \tag{20}$$

where ψ is some non–linear function of total expenditure. In particular, Banks et al. (1997), using 1980-1982 U.K. FES data, found that Engel curves have indeed the form of equation (20), with $\psi(x_h) = \log^2 x_h$. In this latter respect, our results slightly differ from previous findings, since here the second Engel curve increases more rapidly for large levels of expenditure.

Derivatives of basic Engel curves. A final way to interpret the factors is based on the estimation of the derivatives of the basic Engel curves. Indeed, the sign of these functions is strictly connected to whether a category of expenditure should be classified as luxury or necessity. In figure 3 (b-d-f) we show the derivatives of the basic Engel curves $\widetilde{\delta}_r^*(x_h)$, estimated with a local–linear fit as explained in Proposition 1 together with 68% and 90% boostrapped confidence intervals. In agreement with the findings above, the first derivative of the first basic Engel curve is negative for families with income below the median, i.e. the poorest ones, as predicted from the Engel law for necessary goods. The derivative of the second Engel curve captures luxuries being positive for medium–high income households, while the derivative of the third curve is zero for most households indicating a constant Engel curve.

The total expenditure elasticity ϵ_j of good j has a direct connection with the double log model, since for any category of expenditure j we can write (see Deaton and Muellbauer, 1980, p. 17):

$$\log w_{jh} = (\epsilon_j - 1) \log x_h + \nu_{jh}, \quad j = 1, \ldots, J, \; h = 1, \ldots, H, \tag{21}$$

where ν_{jh} is an error term. In our framework, the latent factors are weighted averages of budget shares thus we can think of a model analogous to (21) for the factors themselves:

$$\log f_{rh} = (\rho_r - 1) \log x_h + \upsilon_{rh}, \quad r = 1, \ldots, R, \; h = 1, \ldots, H.$$

Thus, if a factor is supposed to represent necessities, we should expect that the derivative of the log–factor with respect to $\log x_h$ is less than zero, i.e. it has elasticity $\rho_r < 1$ (and $\rho_r > 1$ if it represents luxuries). After rescaling the estimated and identified factor in such a way that $\widetilde{f}_{rh} > 0$, we estimate the average (over households) derivative $\frac{\partial \widetilde{f}_{rh}}{\partial x_h}$, since it has the same sign as $\frac{\partial \log \widetilde{f}_{rh}}{\partial \log x_h}$, being in this case both $\widetilde{f}_{rh}$ and x_h greater than zero. In particular, we estimate average derivatives in a way, using the method proposed by Härdle and Stoker (1989), which being based on kernel density estimates does not require to assume any functional form of the factors. Table 9 displays the estimated average derivatives together with results from the Wald test for zero derivative. The null hypothesis is rejected at the 5% significance level for the first factor when considering only poor households and for the second factor when considering only rich households. This result together with the signs of the derivatives confirm that the first factor captures expenditures for necessities which are more related to low income families, while the second factor captures mainly expenditure for luxuries which are a feature of the behavior of higher income families. The third factor captures goods with income elasticity close to unit, i.e. zero derivative.

8 Conclusions

In this paper, we propose a method to determine the rank of a system of Engel curves for different categories of expenditures expressed in budget shares form. The rank of such a system determines the maximum number of functions of total expenditure, which we call *basic* Engel curves, that drive consumers' behaviour. We frame the problem of finding the rank as the problem of determining the number of latent common factors explaining variations of the system of budget shares. Herein, we identify the maximum number of common factors by means of the criterion proposed by Bai and Ng (2002). The factors can be estimated via approximate principal components and then identified by independent component analysis.

We apply this method to U.K. Family Expenditure Survey annual data. In order to apply factor analysis, we build a large dimension panel of data, in which the budget shares, which are relative to 13 categories of expenditures, of 100 representative households are pooled over different years. The way this dataset is built is based on the method proposed by Kneip (1994). This large dimensional dataset permits us to eschew any assumption of non–correlation among idiosyncratic shocks. The departure from the Gaussian distribution that budget shares display and a hypothesis about the nature of the fundamental drivers of consumption decisions permit us to apply independent component analysis to achieve identification.

Once the common latent factors are identified, we study their properties by means of non–parametric regressions which are consistent estimates of the basic Engel curves. To compare our results with the existing literature we also estimate parametric models for the factors as non–linear functions of total expenditure. Finally, we estimate the first derivatives of the basic Engel curves by applying local–linear regressions and the method proposed by Härdle and Stoker (1989). All results show that the observed system of budget shares is well represented by the sum of a logarithmic, log–quadratic, and constant basic Engel curves, in a form which is consistent with the model suggested by Lewbel (1997). Moreover, the three sources of consumption variation reflect those consumption behaviours typical of expenditures for necessities, luxuries, and unity elasticity goods.

References

Alessi, L., M. Barigozzi, and M. Capasso (2010). Improved penalization for determining the number of factors in approximate static factor models. *Statistics and Probability Letters 80*.

Altman, N. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association 85*(411), 749–759.

Aversi, R., G. Dosi, G. Fagiolo, M. Meacci, and C. Olivetti (1999). Demand dynamics with socially evolving preferences. *Industrial and Corporate Change 8*, 353–468.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*, 191–221.

Banerjee, A. and E. Duflo (2011). *Poor economics: a radical rethinking of the way to fight global poverty*. Public Affairs.

Banks, J., R. Blundell, and A. Lewbel (1997). Quadratic Engel curves and consumer demand. *Review of Economics and Statistics 79*, 527–539.

Belouchrani, A., K. Abed Meraim, J. Cardoso, and E. Moulines (1997). A blind source separation tecnique based on second order statistics. *IEEE Transactions on Signal Processing 45*.

Blundell, R., X. Chen, and D. Kristensen (2007). Semi–non–parametric IV estimation of shape-invariant Engel curves. *Econometrica 75*(6), 1613–1669.

Blundell, R., A. Duncan, and K. Pendakur (1998). Semiparametric estimation and consumer demand. *Journal of Applied Econometrics 13*(5), 435–461.

Blundell, R. W., M. Browning, and I. A. Crawford (2003). Non–parametric Engel curves and revealed preference. *Econometrica 71*(1), 205–240.

Bonhomme, S. and J. Robin (2009). Consistent noisy independent component analysis. *Journal of Econometrics 149*, 12–25.

Cardoso, J. and A. Souloumiac (1993). Blind beamforming for non-Gaussian signals. *IEE Proceedings part F Radar and Signal Processing 140*, 362–362.

Chai, A. and A. Moneta (2010). Retrospectives: Engel curves. *The Journal of Economic Perspectives 24*(1), 225–240.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica 51*, 1305–1324.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing 36*, 287–314.

Deaton, A. and J. Muellbauer (1980). An almost ideal demand system. *American Economic Review 70*, 312–326.

Donald, S. (1997). Inference concerning the number of factors in a multivariate nonparamentric relationship. *Econometrica 65*, 103–132.

Doz, C., D. Giannone, and L. Reichlin (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics 164*(1), 188–205.

Doz, C., D. Giannone, and L. Reichlin (2012). A quasi maximum likelihood approach for large approximate dynamic factor models. *Review of Economics and Statistics 94*(4), 1014–1024.

Engel, E. (1857). Die Productions- und Consumtionsverhältnisse des Königreichs Sachsen. *Bulletin de l'Institut International de la Statistique* (9).

Fagiolo, G., L. Alessi, M. Barigozzi, and M. Capasso (2010). On the distributional properties of household consumption expenditures: The case of Italy. *Empirical Economics 38*.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Annals of Statistics 21*, 196–216.

Fan, J. and I. Gijbels (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics 20*, 2008–2036.

Fan, J. and I. Gijbels (2003). *Local Polynomial Modelling and Its Applications*. Chapam & HallCRC.

Foellmi, R. and J. Zweimüller (2008). Structural change, Engel's consumption cycles and kaldor's facts of economic growth. *Journal of Monetary Economics 55*(7), 1317–1328.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics 82*, 540–554.

Gasser, T. and G. Müller, H (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics 11*, 171–185.

Gorman, W. M. (1981). Some Engel curves. In A. Deaton (Ed.), *Essays in the Theory and Measurements of Consumer Behaviour in Honor of Sir Richard Stone*. Cambridge University Press.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica 52*(3), 681–700.

Härdle, W. (1990). *Applied Non–parametric Regression*, Volume 5. Cambridge Univ Press.

Härdle, W. and T. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association 84*, 986–995.

Hildenbrand, W. (2008). aggregation (theory). In S. N. Durlauf and L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan.

Hyvärinen, A., J. Karhunen, and E. Oja (2001). *Independent Component Analysis*. Wiley.

Hyvärinen, A. and E. Oja (2000). Independent component analysis: Algorithms and applications. *Neural Networks 13*, 411–430.

Jorgensen, DW, L. L. and T. Stoker (1982). The transcendental logarithmic model of aggregate consumer behavior. *Advances in Econometrics*.

Kirman, A. (1992). Whom or what does the representative individual represent? *The Journal of Economic Perspectives 6*(2), 117–136.

Kneip, A. (1994). Non–parametric estimation of common regressors for similar curve data. *The Annals of Statistics 22*, 1386–1427.

Lewbel, A. (1991). The rank of demand systems: Theory and non–parametric estimation. *Econometrica 59*, 711–730.

Lewbel, A. (1997). Consumer demand systems and household equivalence scales. *Handbook of Applied Econometrics: Microeconomics 2*, 167–201.

Lewbel, A. (2001). Demand systems with and without errors. *American Economic Review*, 611–618.

Lewbel, A. (2003). A rational rank four demand system. *Journal of Applied Econometrics 18*(2), 127–135.

Onatski, A. (2010). Determining the number of factors form empirical distribution of eigenvalues. *Review of Economics and Statistics*. forthcoming.

Pasinetti, L. (1981). *Structural change and economic growth: a theoretical essay on the dynamics of the wealth of nations*. Cambridge University Press.

Stock, J. and M. Watson (1989). New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, 351–394.

Stoker, T. (1993). Empirical approaches to the problem of aggregation over individuals. *Journal of Economic Literature 31*, 1827–1874.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya 26*, 359–372.

Witt, U. (1999). Bioeconomics as economics from a darwinian perspective. *Journal of Bioeconomics 1*(1), 19–34.

Witt, U. (2001). Learning to consume–A theory of wants and the growth of demand. *Journal of Evolutionary Economics 11*, 23–36.

A Description of the JADE algorithm

Assume to know the R-dimensional vector of factors $\mathbf{f}_h$, then its cumulant generating function is defined as

$$\mathcal{K}(\boldsymbol{\xi}) = \log \mathrm{E}\left[\exp(\boldsymbol{\xi}'\mathbf{f})\right].$$

We are interested in the fourth–order cumulants which are the the coefficients of the fourth–order terms in the Taylor approximation of $\mathcal{K}(\boldsymbol{\xi})$ in a neighborhood of $\boldsymbol{\xi} = \mathbf{0}$, thus if $\mathrm{E}[\mathbf{f}] = \mathbf{0}$ we have

$$\kappa_{ijk\ell} = \mathrm{E}[f_i f_j f_h f_\ell].$$

There are R^4 fourth order cumulants. All these cumulants can be collected into a single $R^2 \times R^2$ matrix, which in turn has R^2 eigenvectors of size $R^2 \times 1$ and each of them can be transformed into a matrix $\mathcal{V}_i$ containing only $R \times R$. The JADE algorithm look for the $R \times R$ matrix $\widehat{\mathbf{U}}$ such that

$$\widehat{\mathbf{U}} = \arg\min_{\mathbf{V}} \sum_{i=1}^{R^2} \mathrm{off}\left(\mathbf{V}'\mathcal{V}_i\mathbf{V}\right) = \arg\min_{\mathbf{V}} \phi(\widehat{\mathbf{f}}), \qquad \text{(A-1)}$$

where $\mathrm{off}(\mathbf{A})$ takes the off–diagonal elements of the matrix $\mathbf{A}$.

B Technical appendix

B.1 Preliminary results

We first need to prove consistency of the estimated and identified factors $\widetilde{f}_{rh}$.

Lemma 1. *Given assumptions 1-5, the estimated and identified factors $\widetilde{f}_{rh}$ are consistent estimators of the true factors, i.e. for any $h = 1, \dots H$*

$$(\widetilde{f}_{rh} - f_{rh})^2 = O_p\left(\min\left(H^{-1}, J^{-1}\right)\right), \qquad r = 1, \dots R,$$

as $J, H \to \infty$.

Proof. First consider the estimated factors $\widehat{f}^k_{rh}$ as the k largest approximate principal components for a generic number of factors k, i.e. obtained by solving (11). Then the estimated number of factors obtained from (12) is such that (see Theorem 2 in Bai and Ng, 2002):

$$p\text{-}\lim_{J,H\to\infty} \widehat{R} = R.$$

The estimated factors are then the $\widehat{R}$ largest principal components: $\widehat{f}_{rh} = \frac{1}{J}\sum_{j=1}^{J} w_{jh}\widehat{a}_{jr}$, where $\widehat{a}_{jr}$ is the entry j of the normalized eigenvector corresponding to the r-th eigenvalue of the sample covariance matrix of $\mathbf{w}_h$. From a corollary of Theorem 1 in Bai and Ng (2002) we have

$$\left|\left|\widehat{\mathbf{f}}_h - \mathbf{U}\mathbf{f}_h\right|\right|^2 = O_p\left(\min\left(J^{-1}, H^{-1}\right)\right), \text{ for any } h = 1, \dots, H, \tag{B-1}$$

where $\mathbf{U}$ is a matrix of rank r.

If we assume statistical independence among the R components of the factors $\mathbf{f}_h$ (see assumptions 4 and 5) then $\mathbf{U}$ is uniquely identifiable. For example from JADE we obtain an estimate $\widehat{\mathbf{U}}$ such that $\widehat{\mathbf{U}}'\widehat{\mathbf{f}}_h$ has R statistically independent components. Moreover, since from (A-1) and the fact that sample cumulants are continuous function of the factors, and by virtue of (B-1), we have

$$(\phi(\widehat{\mathbf{f}}) - \phi(\mathbf{U}\mathbf{f}))^2 = O_p\left(\min\left(J^{-1}, H^{-1}\right)\right),$$

which implies

$$\left|\left|\widehat{\mathbf{U}}_{\widehat{\mathbf{f}}} - \widehat{\mathbf{U}}_{\mathbf{U}\mathbf{f}}\right|\right|^2 = O_p\left(\min\left(J^{-1}, H^{-1}\right)\right). \tag{B-2}$$

where $\widehat{\mathbf{U}}_{\widehat{\mathbf{f}}}$ is the maximizer of (A-1) when using the fourth–order cumulants of $\widehat{\mathbf{f}}$ and analogously we define $\widehat{\mathbf{U}}_{\mathbf{U}\mathbf{f}}$.

Since for any vector $\mathbf{x}$, JADE determines $\widehat{\mathbf{U}}_{\mathbf{x}}$ in order to make the components of the vector $\widehat{\mathbf{U}}'_{\mathbf{x}}\mathbf{x}$ statistically independent, then $\widehat{\mathbf{U}}'_{\mathbf{U}\mathbf{f}}\mathbf{U}\mathbf{f}_h$ has statistically independent components. Given that the ICA problem has a unique solution up to a sign, a scale, and a permutation, and given that by assumption $\mathbf{f}_h$ has already independent components, then we must have $\widehat{\mathbf{U}}'_{\mathbf{U}\mathbf{f}}\mathbf{U} = \mathbf{I}_r$. Indeed we can fix the sign, scale, and permutation indeterminacy by adding assumptions on the true factors as described in the main text.

By multiplying both terms in (B-1) by $\widehat{\mathbf{U}}_{\widehat{\mathbf{f}}}$ we have

$$\left|\left|\widehat{\mathbf{U}}_{\widehat{\mathbf{f}}}\widehat{\mathbf{f}}_h - \widehat{\mathbf{U}}_{\widehat{\mathbf{f}}}\mathbf{U}\mathbf{f}_h\right|\right|^2 \leq \left|\left|\widehat{\mathbf{U}}_{\widehat{\mathbf{f}}}\widehat{\mathbf{f}}_h - \widehat{\mathbf{U}}_{\mathbf{U}\mathbf{f}}\mathbf{U}\mathbf{f}_h\right|\right|^2 + \left|\left|\widehat{\mathbf{U}}_{\mathbf{U}\mathbf{f}}\mathbf{U}\mathbf{f}_h - \widehat{\mathbf{U}}_{\widehat{\mathbf{f}}}\mathbf{U}\mathbf{f}_h\right|\right|^2, \; h = 1, \dots, H,$$

From (B-2) we have that the second term on the right–hand–side is $O_p\left(\min\left(J^{-1}, H^{-1}\right)\right)$. Moreover, from (B-1) also the term on the left–hand–side is $O_p\left(\min\left(J^{-1}, H^{-1}\right)\right)$, therefore also the first term on the right–hand–side must be $O_p\left(\min\left(J^{-1}, H^{-1}\right)\right)$. If we define $\widetilde{\mathbf{f}}_h = \widehat{\mathbf{U}}_{\widehat{\mathbf{f}}}\widehat{\mathbf{f}}_h$ and recalling that $\widehat{\mathbf{U}}'_{\mathbf{U}\mathbf{f}}\mathbf{U} = \mathbf{I}_r$, this latter term becomes

$$\left|\left|\widehat{\mathbf{U}}_{\widehat{\mathbf{f}}}\widehat{\mathbf{f}}_h - \widehat{\mathbf{U}}_{\mathbf{U}\mathbf{f}}\mathbf{U}\mathbf{f}_h\right|\right|^2 = \left|\left|\widetilde{\mathbf{f}}_h - \mathbf{f}_h\right|\right|^2 = O_p\left(\min\left(J^{-1}, H^{-1}\right)\right), \text{ for any } h = 1, \dots, H,$$

or equivalently

$$\left|\widetilde{f}_{rh} - f_{rh}\right|^2 = O_p\left(\min\left(J^{-1}, H^{-1}\right)\right), \text{ for any } h = 1, \dots, H, \; r = 1, \dots R, \tag{B-3}$$

which proves Lemma 1. □

We make the following assumption on the kernel function and the bandwidth.

Assumption K. The kernel function is such that:

$$\begin{aligned}\int_{-1}^{1} \mathrm{K}_{b_H}(u)\mathrm{d}u &= 1,\\ \int_{-1}^{1} u\mathrm{K}_{b_H}(u)\mathrm{d}u &= 0,\\ \int_{-1}^{1} u^2\mathrm{K}_{b_H}(u)\mathrm{d}u &< \infty.\end{aligned}$$

The bandwidth b_H is such that $b_H \to 0$, $Hb_H \to \infty$, and $Hb_H^2 \to 0$ as $H \to \infty$. Moreover it satisfies $b_H = O(H^{-d})$ with $d < 1$ when $H \to \infty$.

B.2 Proof of Proposition 1

For any x_h and any $r = 1, \ldots, R$ we have the following local linear estimators for the basic Engel curves $\widetilde{\gamma}_r^*(x_h)$ and their first–derivative $\widetilde{\delta}_r^*(x_h)$:

$$\begin{pmatrix} \widetilde{\gamma}_r^*(x_h) \\ \widetilde{\delta}_r^*(x_h) \end{pmatrix} = \arg\max_{\gamma_r,\delta_r} \sum_{k=1}^{H} \left[\widetilde{f}_{rk} - \gamma_r - \delta_r(x_k - x_h)\right]^2 \mathrm{K}_{b_H}(x_k - x_h), \quad r = 1, \ldots, R.$$

If we define

$$\mathbf{Z}_k(x_h) = \begin{pmatrix} 1 \\ x_k - x_h \end{pmatrix}$$

the closed form expression for the estimators is given by

$$\begin{pmatrix} \widetilde{\gamma}_r^*(x_h) \\ \widetilde{\delta}_r^*(x_h) \end{pmatrix} = \left(\sum_{k=1}^{H} \mathbf{Z}_k(x_h)\mathbf{Z}_k'(x_h)\mathrm{K}_{b_H}(x_k - x_h)\right)^{-1} \left(\sum_{k=1}^{H} \mathbf{Z}_k(x_h)\widetilde{f}_{rk}\mathrm{K}_{b_H}(x_k - x_h)\right).$$

With respect to the main text and assumption 6 we consider here the case $\mathrm{E}[z_{rh}^2] = \sigma^2$ for any h. The generalization to the heteroskedastic case is straightforward. The, from e.g. Fan and Gijbels (2003) and Härdle (1990) we know that $\widetilde{\gamma}_r^*(x_h)$ is an estimator of $\widetilde{\gamma}_r(x_h) = \mathrm{E}[\widetilde{f}_{rh}|x_h]$ such that, for any x_h,

$$|\widetilde{\gamma}_r^*(x_h) - \widetilde{\gamma}_r(x_h)|^2 = O_p(H^{-1}b_H^{-1}) + \kappa B O_p(H^{-1}b_H^{-1}) + \sigma^2 C(1 + 2S) O_p(H^{-1}b_H^{-1}). \quad \text{(B-4)}$$

In the expression above κ depends on the second derivative of $\widetilde{\gamma}_r(x_h)$, σ^2 is the variance of the errors, while

$$
\begin{aligned}
B &= \int_{-1}^{1} u^2 \mathrm{K}_{b_H}(u)\mathrm{d}u, \\
C &= \int_{-1}^{1} \mathrm{K}_{b_H}^2(u)\mathrm{d}u, \\
S &= \lim_{k\to\infty} \sum_{j=1}^{k} \rho(j).
\end{aligned}
$$

where $\rho(j) \sim |j|^{-\zeta}$. The third term in (B-4) is due to the correlation in the errors. Since by assumption 6, $\zeta > 1$, we have $S < \infty$, and this term is also $O_p(H^{-1}b_H^{-1})$. Therefore,

$$
|\widetilde{\gamma}_r^*(x_h) - \widetilde{\gamma}_r(x_h)|^2 = O_p(H^{-1}b_H^{-1}) + O_p(b_H^2 H^{-1} b_H^{-1}). \tag{B-5}
$$

Now consider the following decomposition, which holds for any $r = 1, \dots R$ and any $h = 1, \dots H$

$$
\begin{aligned}
|\widetilde{\gamma}_r^*(x_h) - g_r(x_h)|^2 &= |\widetilde{\gamma}_r^*(x_h) - \widetilde{\gamma}_r(x_h) + \widetilde{\gamma}_r(x_h) - g_r(x_h)|^2 \leq \\
&\leq |\widetilde{\gamma}_r^*(x_h) - \widetilde{\gamma}_r(x_h)|^2 + |\widetilde{\gamma}_r(x_h) - g_r(x_h)|^2.
\end{aligned}
$$

Given (B-5), the first term in the last inequality is $O_p(H^{-1}b_H^{-1}) + O_p(b_H H^{-1})$, while the second term can be written as

$$
\left|\mathrm{E}\big[\widetilde{f}_{rh}|x_h\big] - \mathrm{E}\big[f_{rh}|x_h\big]\right|^2 = \left|\mathrm{E}\big[(\widetilde{f}_{rh} - f_{rh})|x_h\big]\right|^2 \leq \mathrm{E}\left[\left|\widetilde{f}_{rh} - f_{rh}\right|^2 |x_h\right] = O_p\left(\min\left(J^{-1}, H^{-1}\right)\right).
$$

where the last equality is given by Lemma 1. Therefore,

$$
|\widetilde{\gamma}_r^*(x_h) - g_r(x_h)|^2 \leq O_p(H^{-1}b_H^{-1}) + O_p(b_H H^{-1}) + O_p\left(\min\left(J^{-1}, H^{-1}\right)\right). \tag{B-6}
$$

Since when $H \to \infty$ assumption K implies $b_H \to 0$, $Hb_H \to \infty$, and $b_H H^{-1} \to 0$, we have

$$
p\text{-}\lim_{J,H\to\infty} |\widetilde{\gamma}_r^*(x_h) - g_r(x_h)|^2 = 0,
$$

which proves Proposition 1, the rate of convergence being given by (B-6). An analogous argument allows us to prove consistency of the first derivative, i.e.

$$
p\text{-}\lim_{J,H\to\infty} |\widetilde{\delta}_r^*(x_h) - g_r'(x_h)|^2 = 0.
$$

□

B.3 Proof of Proposition 2

The proof is based on the same arguments as the proof of Proposition 1. Consider the equation

$$\widetilde{f}_{rh} = \widetilde{\alpha}_r + \widetilde{\beta}_r m(x_h) + \widetilde{z}_{rh}, \ r = 1, \dots, R; \ h = 1, \dots, H, \tag{B-7}$$

then define

$$\mathcal{X} = (\mathbf{1}_H, m(\mathbf{x})), \qquad \widetilde{\boldsymbol{\theta}}_r = \begin{pmatrix} \widetilde{\alpha}_r \\ \widetilde{\beta}_r \end{pmatrix}, \quad r = 1, \dots, R,$$

where $\mathbf{1}_H$ is an H-dimensional column vector of ones and $\mathbf{x} = (x_1 \dots x_H)'$. The least squares estimator of (B-7)

$$\widetilde{\boldsymbol{\theta}}^*_r = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\widetilde{\mathbf{f}}_r,$$

is such that (see e.g. Gourieroux et al., 1984)

$$p\text{-}\lim_{H\to\infty} |\widetilde{\boldsymbol{\theta}}^*_r - \widetilde{\boldsymbol{\theta}}_r| = 0, \quad r = 1, \dots, R, \tag{B-8}$$

with rate $H^{-1/2}$. Now consider, for any $r = 1, \dots, R$,

$$|\widetilde{\boldsymbol{\theta}}^*_r - \boldsymbol{\theta}_r| = |\widetilde{\boldsymbol{\theta}}^*_r - \widetilde{\boldsymbol{\theta}}_r + \widetilde{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r| \leq |\widetilde{\boldsymbol{\theta}}^*_r - \widetilde{\boldsymbol{\theta}}_r| + |\widetilde{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r| \tag{B-9}$$

the first term on the right–hand–side is $O_p(H^{-1/2})$. If we multiply the second term by $\mathcal{X}$ we have

$$\mathcal{X}|\widetilde{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r| = |\mathcal{X}\widetilde{\boldsymbol{\theta}}_r - \mathcal{X}\boldsymbol{\theta}_r| = \mathrm{E}\big[\widetilde{\mathbf{f}}_r - \mathbf{f}_r|\mathcal{X}\big] = O_p\left(\min\left(J^{-1/2}, H^{-1/2}\right)\right), \tag{B-10}$$

by Lemma 1. By combining (B-8), (B-9), and (B-10) we get the required result. □

Tables and figures

Table 1: Building the dataset

Step 1 *Deflation*:
Let (X_{it}, Y^g_{it}) be the original dataset, where X_{it} is total expenditure and Y^g_{it} is expenditure on category $g = 1, \dots, 13$. The subscript refers to household i surveyed at time (year) t. At year $t = 1, \dots, T$ ($T = 10$ in the time wave analysed here), the number of surveyed households is I_t. Let P_t be the retail price index at year t. Let P^g_{it} be the sub-index of the retail price index corresponding to g at year t. Consider the data (X^*_{it}, W^g_{it}), where $X^*_{it} = X_{it}/P_{i_t}$ and $W^g_{it} = Y^g_{it}/(P^g_{i_t} X^*_{it})$, for each good $g = 1, \dots, 13$, each household $i_t = 1, \dots, I_t$, and for each $t = 1, \dots, T$. We omit the index t when we refer to a fixed year.

Step 2 *Remove outliers*:
For each year $t = 1, \dots, T$ calculate the sample mean and standard deviation of X^*_{it}. Trim (for each t) data whose X^*_{it} values lie outside three standard deviations from the mean. Let min(X^*_{it}) be the smallest value of X^*_{it} for each year t, and max(X^*_{it}) the highest value of X^*_{it} for each year t. For each t remove data whose X^*_{it} values lie beneath the maximum (over time) of min(X^*_{it}) and beyond the minimum of max(X^*_{it}).

Step 3 *Segmenting total expenditure*:
Consider the percentiles of X^*_{it}, corresponding to 100 values $k_{1,t}, \dots, k_{100,t}$ of X^*_{it} for each t. For each percentile, that is for each $h = 1, \dots, 100$, take the average over time, i.e. $\kappa_h = \frac{1}{T} \sum_{t=1}^{T} k_{ht}$. Let κ_0 be the lowest value of X^*_{it} (for each i and t). We let $[\kappa_0, \kappa_1]$, $[\kappa_1, \kappa_2] \dots$, $[\kappa_{99}, \kappa_{100}]$ determine 100 intervals of total expenditure.

Step 4 *Averaging budget shares within intervals*:
Separately for each $t = 1, \dots, T$, each $g = 1, \dots, 13$ and each $h = 1, \dots, 100$, let

$$W^{g*}_{ht} = \frac{\sum_{i=1}^{I_t} W^g_{it} \mathcal{I}_{[\kappa_{h-1}, \kappa_h]}(X^*_{it})}{\sum_{i=1}^{I_t} \mathcal{I}_{[\kappa_{h-1}, \kappa_h]}(X^*_{it})},$$

where $\mathcal{I}_{[A]}(x) = 1$ when $x \in A$ and 0 otherwise. This corresponds to taking average of budget shares within each interval. We thus have 100 representative families with $13 \cdot T$ different budget allocations. Let $J = 13 \cdot T$.

Step 5 *New dataset*:
Let the new dataset be (x_h, w_{hj}), with $x_h = (\kappa_h - \kappa_{h-1})/2$, and $w_{h1} = W^{1*}_{h1}; w_{h2} = W^{2*}_{h1}; \dots; w_{h13} = W^{13*}_{h1}; w_{h14} = W^{1*}_{h2}; \dots; w_{hJ} = W^{13*}_{hT}$ (for each $h = 1, \dots, 100$).

Table 2: Determining the number of factors in single years.

	household members			
years	**1**	**2**	**2-3**	**2-4**
1997	1	5	1	1
1998	4	4	4	3
1999	3	2	2	1
2000	3	1	1	1
2001	4	3	2	2
2002	1	2	1	1
2003	1	1	1	4
2004	3	2	1	4
2005	1	2	2	2
2006	1	2	3	2
average	2.2	2.4	1.8	2.1

Number of factors based on the LDU decomposition (see Lewbel, 1991) obtained at 5% significance level under the null–distribution which is $\chi^2_{(13-R_t)}$.

Table 3: Comparison of factor estimates.

	household members															
	1				**2**				**2-3**				**2-4**			
	num. factors				**num. factors**				**num. factors**				**num. factors**			
years	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
1997	0.94	0.89	0.87	0.85	0.95	0.93	0.91	0.89	0.96	0.94	0.92	0.90	0.96	0.94	0.93	0.91
1998	0.93	0.88	0.85	0.83	0.95	0.92	0.90	0.88	0.96	0.94	0.92	0.91	0.97	0.95	0.93	0.92
1999	0.91	0.84	0.82	0.79	0.94	0.89	0.88	0.86	0.95	0.92	0.91	0.89	0.97	0.93	0.92	0.91
2000	0.94	0.89	0.85	0.83	0.96	0.91	0.89	0.87	0.97	0.93	0.91	0.90	0.96	0.93	0.91	0.90
2001	0.94	0.90	0.87	0.84	0.95	0.91	0.89	0.87	0.96	0.92	0.90	0.88	0.96	0.92	0.91	0.89
2002	0.94	0.89	0.86	0.84	0.96	0.92	0.89	0.87	0.96	0.93	0.91	0.89	0.97	0.93	0.92	0.90
2003	0.95	0.89	0.85	0.83	0.94	0.90	0.87	0.86	0.96	0.89	0.87	0.87	0.97	0.90	0.89	0.88
2004	0.94	0.87	0.84	0.83	0.94	0.87	0.85	0.84	0.96	0.91	0.90	0.88	0.96	0.92	0.91	0.90
2005	0.93	0.87	0.84	0.82	0.94	0.88	0.86	0.84	0.96	0.92	0.90	0.89	0.97	0.92	0.91	0.89
2006	0.93	0.88	0.84	0.82	0.95	0.91	0.88	0.87	0.97	0.92	0.90	0.89	0.98	0.93	0.92	0.90
average	0.94	0.88	0.85	0.83	0.95	0.90	0.88	0.86	0.96	0.92	0.90	0.89	0.97	0.93	0.91	0.90

Trace statistics τ_t when regressing the factors estimated on the pooled dataset on the factors estimated on single blocks indicated by the years.

Table 4: Average budget shares.

	household members											
	1			**2**			**2-3**			**2-4**		
	all	**poor**	**rich**	**all**	**poor**	**rich**	**all**	**poor**	**rich**	**all**	**poor**	**rich**
Housing	0.23	0.22	0.24	0.19	0.19	0.18	0.18	0.19	0.18	0.18	0.19	0.18
Fuel, light and power	0.07	0.10	0.04	0.05	0.07	0.03	0.05	0.06	0.03	0.04	0.06	0.03
Food	0.19	0.24	0.14	0.19	0.23	0.14	0.19	0.23	0.15	0.19	0.23	0.15
Alcoholic drinks	0.04	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Tobacco	0.03	0.04	0.02	0.02	0.03	0.01	0.02	0.03	0.01	0.02	0.03	0.01
Clothing and footwear	0.03	0.02	0.04	0.04	0.03	0.05	0.04	0.04	0.05	0.05	0.04	0.05
Household goods	0.07	0.06	0.07	0.08	0.07	0.08	0.07	0.07	0.08	0.07	0.07	0.08
Household services	0.07	0.08	0.07	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.05	0.06
Personal goods and services	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Motoring	0.08	0.04	0.12	0.12	0.10	0.15	0.13	0.10	0.15	0.13	0.10	0.15
Travels	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Leisure goods	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Leisure services	0.11	0.09	0.12	0.12	0.10	0.14	0.12	0.10	0.14	0.12	0.09	0.14

Averages are computed over the 10 years period 1997–2006; **poor**: households with total expenditure below median; **rich**: households with total expenditure above median.

Table 5: Determining the number of factors and their explained variance on pooled data.

	household members											
	1			**2**			**2-3**			**2-4**		
criterion	**77–86**	**87–96**	**97–06**	**77–86**	**87–96**	**97–06**	**77–86**	**87–96**	**97–06**	**77–86**	**87–96**	**97–06**
BN	12	3	2	4	3	2	5	3	2	5	3	2
ABC	2	3	2	4	3	2	4	3	2	4	4	4
O	2	2	2	2	3	2	2	3	2	2	3	2
CC	4	3	4	4	3	3	4	3	4	4	3	3
average	5	2.75	2.5	3.5	3	2.25	3.75	3	2.5	3.5	3.25	2.75
EV												
Factor 1	0.41	0.52	0.55	0.56	0.59	0.63	0.64	0.65	0.68	0.70	0.69	0.69
Factor 2	0.18	0.19	0.06	0.09	0.12	0.05	0.09	0.12	0.07	0.08	0.12	0.08
Factor 3	0.04	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01

BN: Bai and Ng (2002) criterion. **ABC**: Alessi et al. (2010) criterion. **O**: Onatski (2010) criterion. **CC**: number of factors based on the canonical correlations obtained at 10% significance level under the null–distribution which is $\chi^2_{(J/2-R)(J/2-R)}$. **EV**: variance explained by each factor computed with respect to total variance.

Table 6: Factor loadings.

	Average Loading		
	Factor 1	**Factor 2**	**Factor 3**
Housing	-0.21	-0.28	0.38
Fuel, light and power	0.34	-0.23	0.21
Food	0.70	-0.65	0.61
Alcoholic drinks	-0.07	-0.02	0.03
Tobacco	0.16	-0.15	0.14
Clothing and footwear	-0.07	0.09	-0.08
Household goods	-0.06	0.13	-0.16
Household services	0.04	0.04	0.00
Personal goods and services	-0.02	0.03	-0.02
Motoring	-0.50	0.33	-0.40
Travels	0.00	0.06	-0.07
Leisure goods	-0.03	0.06	-0.08
Leisure services	-0.24	0.52	-0.47

Average loadings of the identified factors $\widetilde{\mathbf{f}}$ are computed over the 10 years period 1997-2006, the scale being fixed such that $\widetilde{\mathbf{A}}'\widetilde{\mathbf{A}}/J = \mathbf{I}_r$. Results refer to households with 2 to 4 members. Loadings for each year are available in the complementary appendix.

Table 7: Parametric estimates of basic Engel curves, goodness–of–fit.

				adj.R^2					
		all			**poor**			**rich**	
Functional form	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**
$\alpha_r + \beta_r x_h$	0.18	0.50	0.27	0.85	0.00	0.06	0.76	0.52	0.22
$\alpha_r + \beta_r x_h^2$	0.04	0.56	0.27	0.71	0.00	0.04	0.73	0.51	0.18
$\alpha_r + \beta_r x_h^{-1}$	0.78	0.08	0.06	0.91	0.13	0.04	0.69	0.48	0.26
$\alpha_r + \beta_r x_h^{-2}$	0.70	0.00	0.01	0.69	0.23	0.00	0.63	0.43	0.26
$\alpha_r + \beta_r \log x_h$	0.50	0.30	0.17	0.95	0.05	0.06	0.74	0.51	0.25
$\alpha_r + \beta_r \log^2 x_h$	0.43	0.34	0.20	0.94	0.03	0.06	0.75	0.51	0.24
$\alpha_r + \beta_r x_h \log x_h$	0.15	0.52	0.28	0.83	0.00	0.05	0.75	0.52	0.21

Adjusted R^2 coefficient for the least squares regressions of the identified factors $\widetilde{f}_{rh}$ on selected functions of total expenditure; **poor**: households with total expenditure below median; **rich**: households with total expenditure above median. Results refer to households with 2 to 4 members.

Table 8: Parametric estimates of basic Engel curves, coefficients.

	all		**poor**		**rich**	
Functional form	$\widetilde{\beta}_r^*$	**adj.R^2**	$\widetilde{\beta}_r^*$	**adj.R^2**	$\widetilde{\beta}_r^*$	**adj.R^2**
$\widetilde{f}_{1h} = \alpha_1 + \beta_1 \log x_h$	-1.18^{*} (0.92)	0.50	-2.89^{***} (1.20)	0.95	0.86 (0.98)	0.74
$\widetilde{f}_{1h} = \alpha_1 + \beta_1 x_h^{-1}$	1.50^{**} (0.89)	0.78	5.26^{***} (2.10)	0.91	-2.58 (3.03)	0.69
$\widetilde{f}_{2h} = \alpha_2 + \beta_2 x_h^2$	0.16^{*} (0.10)	0.56	-0.01 (0.09)	0.00	0.11^{**} (0.06)	0.51
$\widetilde{f}_{2h} = \alpha_2 + \beta_2 x_h \log x_h$	0.44^{*} (0.29)	0.56	-0.03 (0.25)	0.00	0.35^{**} (0.19)	0.52

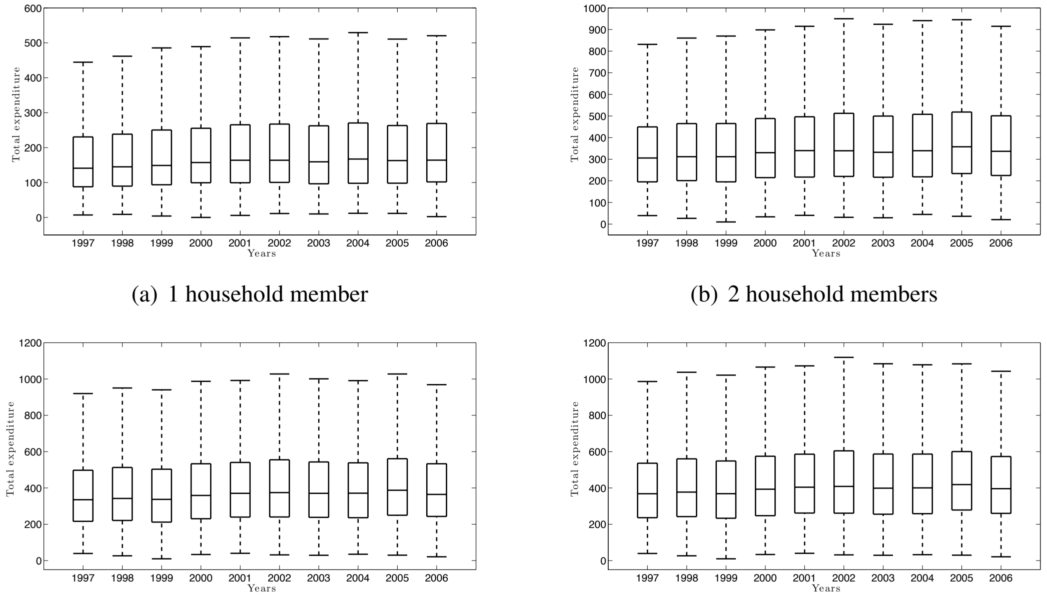
Estimates of the slope coefficient for the regression of the identified factors $\widetilde{f}_{rh}$ on selected functions of total expenditure; standard errors in parenthesis are computed by re–estimating the factors and the Engel curves using 1000 bootstrapped samples of budget share: * significant at 10%; ** significant at 5%; *** significant at 1%; **poor**: households with total expenditure below median; **rich**: households with total expenditure above median. Estimates for the intercept α_r are available upon request. Results refer to households with 2 to 4 members.

Table 9: Average derivatives.

	all			poor			rich		
	Factor 1	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**
Derivative	-0.26	0.32	-0.19	-0.53	0.00	0.14	0.11	0.32	-0.29
Wald statistic	2.35 (0.13)	8.07 (0.00)	2.71 (0.10)	16.41 (0.00)	0.00 (0.99)	3.32 (0.07)	17.00 (0.00)	7.72 (0.01)	7.03 (0.01)

Derivatives averaged across total expenditure estimated using Härdle and Stoker (1989) method; Wald statistics under the null hypothesis of zero average derivative and computed with standard errors obtained with 1000 bootstrap replications (p-values in parenthesis); **poor**: households with total expenditure below median; **rich**: households with total expenditure above median. Results refer to households with 2 to 4 members.

Figure 1: Distribution of total expenditure.



(a) 1 household member

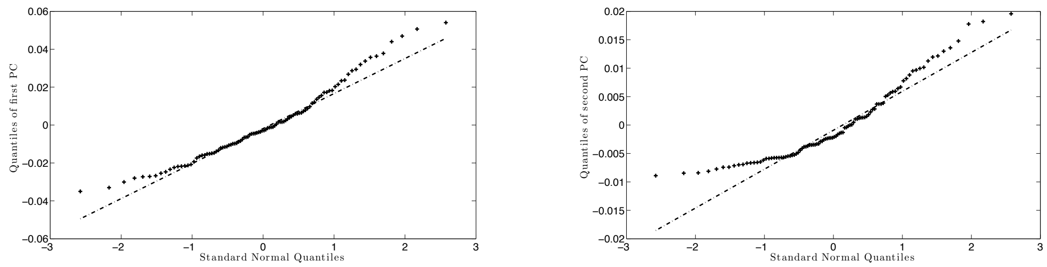
(b) 2 household members

(c) 2–3 household members

(d) 2–4 household members

Box–plots for total expenditure, showing median, 25th, 75th percentiles, minumum and maximum of the distribution.

Figure 2: Non–Gaussianity of the factors.

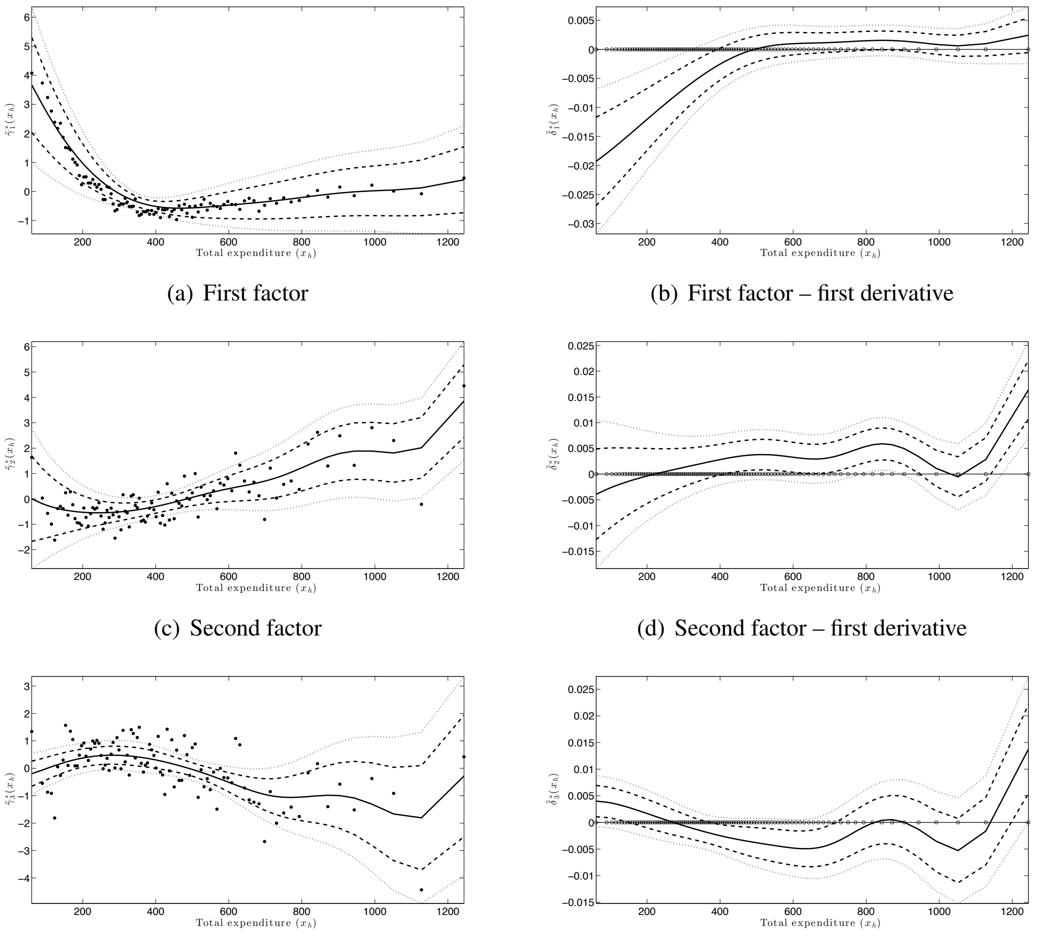


(a) First factor

(b) Second factor

Quantiles of the two largest principal components, i.e. of the estimated factors $\widehat{f}_{rh}$ *vs.* quantiles of a standard Gaussian distribution. Results refer to households with 2 to 4 members and the period 1997–2006.

Figure 3: Estimated basic Engel curves and their first derivatives.



(a) First factor

(b) First factor – first derivative

(c) Second factor

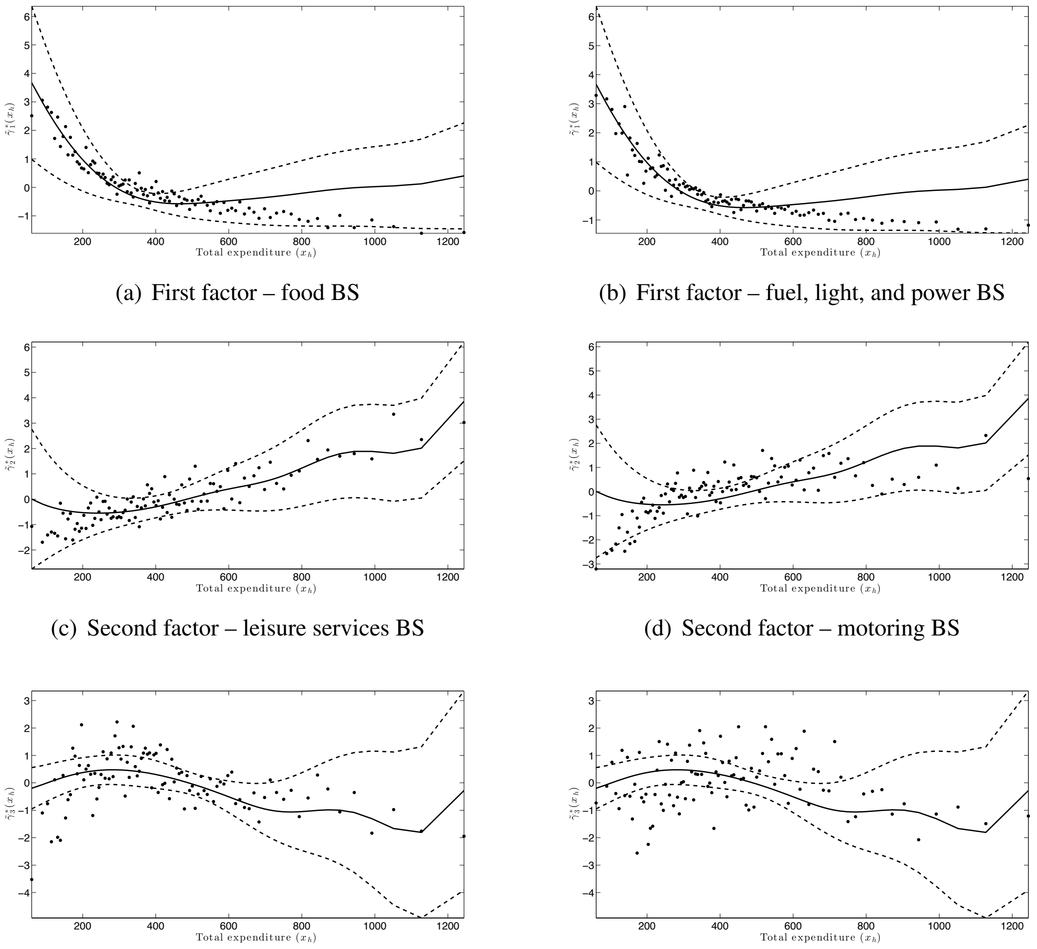
(d) Second factor – first derivative

(e) Third factor

(f) Third factor – first derivative

Solid line: local linear non–parametric estimates of basic Engel curves $\widetilde{\gamma}_r^*(x_h)$ (left column) and their first derivatives $\widetilde{\delta}_r^*(x_h)$ (right column); dashed line: 68% confidence intervals; dotted line: 90% confidence intervals; circles: values taken by the factors $\widetilde{f}_{rh}$ (left column). Confidence intervals are obtained with 1000 bootstrap replications. In this graph factors are re–scaled to have zero mean. Results refer to households with 2 to 4 members and the period 1997–2006.

Figure 4: Interpreting the basic Engel curves.



(a) First factor – food BS

(b) First factor – fuel, light, and power BS

(c) Second factor – leisure services BS

(d) Second factor – motoring BS

(e) Third factor – housing BS

(f) Third factor – alcoholic drinks BS

Circles: budget shares w_{jh} of selected goods in 2006; solid lines: estimated non–parametric basic Engel curves $\widetilde{\gamma}_r^*(x_h)$; dashed lines: 90% confidence intervals obtained with 1000 bootstrap replications. In this graph the non–parametric fits have mean zero and standard deviation one and budget shares are rescaled accordingly. Results refer to households with 2 to 4 members.